part structure of EPA's aquatic life criteria, on which the proposed Idaho criteria are based, with CMC to protect against short-term effects of exposures to criteria chemicals, and a CCC to protect against long or indefinite term exposures, the protectiveness of the CMCs were evaluated against data on effects in short-term exposures (≤ 96 hours) and CCCs were evaluated against data on effects in longer-term exposures.

In most instances, direct testing evidence for the listed salmon species was not available, and test data obtained with other fish species was used as surrogate estimates of potential effects to listed salmon. Steelhead were an exception, since they and rainbow trout are different forms of the same species (Behnke and Tomelleri 2002; Quinn 2005). In most cases, rainbow trout data were available since rainbow trout are commonly tested in ecotoxicology. Rainbow trout are often used as a surrogate for all listed *Oncorhynchus*, using geometric means. At least with several metals, rainbow trout are probably similar in sensitivity to Chinook salmon and probably considerably more sensitive than sockeye salmon. Few direct data with sockeye salmon were located, which may be related to Chapman's (1975) recommendation against testing sockeye salmon following his observations that they were much less sensitive to metals than were Chinook or coho salmon or rainbow/steelheads (Chapman 1975).

In addition to Idaho's aquatic life criteria, EPA has also approved Idaho criteria designed to protect human health from recreational, fish consumption, and drinking water uses which are also applicable to the waters in the action area. In practice, when multiple criteria are applicable to the same water body, the most stringent criteria will drive discharge limits and other pollution management efforts (IDEQ 2007a; subsection 70.1, "Applicability of standards, multiple criteria"). For our analysis, if review of the aquatic life CCC indicated that adverse effects to listed species or their habitats were likely, then we reviewed the human health-based ambient water quality criteria concentrations for the same substance to see if the human-health concentrations would be protective of the listed steelhead and salmon.

2.4.1. Evaluation of issues that are common to multiple aquatic life criteria

All criteria being evaluated as part of this action were developed by EPA following EPA's guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses. For short, these are referred as the "Guidelines" (Stephan *et al*. 1985). Thus it is important to consider the structure of the Guidelines in regard to protection of listed salmon, steelhead, and their critical habitats to evaluate whether criteria derived following them would likely be protective.

The EPA's Guidelines for criteria development represent the best judgments of a committee of EPA scientists as of the mid-1980s. As the title states, the objectives of the criteria development was the "protection of aquatic organisms and their uses." Because the Guidelines are quite detailed and have much explicit guidance, their use has tended to make criteria documents (the supporting documents prepared by EPA in deriving national recommended water quality criteria) objective, transparent, and reproducible. However, the Guidelines recognize that ecotoxicology and criteria derivation cannot be reduced to a series of decision rules, and many judgments are required to produce an individual criteria document. Because the Guidelines are fundamental to criteria, they are fundamental to the evaluation of the protectiveness of criteria for ESA-listed species and habitats. The fundamental assumptions and procedures in the Guidelines are inherent to their degree of protectiveness for listed salmon and steelhead. Thus some of key criteria derivation steps are briefly described here and the underlying assumptions are critically examined.

The Guidelines include some fundamental assumptions:

- Effects which occur on a species in appropriate laboratory tests will generally occur on the same species in comparable field situations.

- For a given substance, if average species sensitivities are rank ordered, the species sensitivity distributes itself in a rather consistent way for most chemicals. Thus, each species tested is not representative of any other species but is one estimate of the general species sensitivity (i.e. a point along the distribution).

- The goal of aquatic life criteria is to protect aquatic communities and socially valued species within those communities. Aquatic organisms may have ecologically redundant functions in communities. The loss of some species might not be important if other species would fill the same ecological function. Thus it is not necessary to protect all of the species all of the time.

- If 95% of the species in acceptable datasets were protected, that would be sufficient to protect aquatic ecosystems in general. In the ecological risk assessment literature, this is often referred to as the 5^{th} percentile of a species sensitivity distribution (SSD) or shortened to the HC5 approach, for the hazardous chemical concentration adversely affecting no more than 5% of the species in a natural community.

- To estimate a criterion protective of 95% of the species, it is acceptable to extrapolate from compilations of severely toxic effects from short-term, “acute” tests to less severe effects in long-term, “chronic” exposures.

- If one or more water quality characteristics such as temperature, pH, or water hardness affect the acute toxicity of a substance in a predictable way, then the acute criterion for that substance should be expressed as a function of that characteristic. It is acceptable to assume that toxicity relationships established with short-term exposure data, such as those between water-hardness and metals toxicity, would be the same in long-term exposures. Thus acute-toxicity and hardness or other relations may be applied equally to chronic criteria (Stephan *et al.* 1985; Stephan 1985; Stephan 2002)

Relying on these assumptions, the EPA Guidelines are derived with the following general steps (Stephan *et al.* 1985):

- First, datasets of acute (short-term) responses of aquatic organisms to the substance of interest are compiled and screened for data sufficiency, relevance and quality.

- If a water quality characteristic is considered to affect the toxicity of the substance, then a relation is developed and the acute data are normalized to a common water condition. For example, with several metals, hardness-toxicity regressions were developed and used to adjust acute toxicity values to a common hardness of 50 mg/L.

- The adjusted acute data are averaged to obtain species mean acute values (SMAVs), and SMAVs are averaged to obtain genus mean acute values (GMAVs). The GMAVs are rank ordered, and value close to the 5^{th} percentile most sensitive genus is calculated, called the final acute value (FAV). The FAV is divided by 2 to extrapolate from a lethal concentration for sensitive taxa to a concentration expected to kill few sensitive taxa. The FAV/2 value becomes the CMC, which is commonly referred to as the acute criterion.

  [In this procedure, if multiple values for a species were available, with differing sensitivities, a geometric mean of all values was taken to calculate the SMAV. If different SMAVs were available, a geometric mean was similarly calculated. For example, with EPA's 1984 copper criteria, the SMAVs for Chinook, Coho and Sockeye salmon were calculated as 42, 70, and 233 µg/L, and a GMAV of 89 µg/L was calculated to represent all *Oncorhynchus*. In that era, steelhead and rainbow trout were considered in a different genus, *Salmo*.]

- Chronic (long-term) data are compiled, and acute-to-chronic ratios (ACRs) are calculated for at least 3 species. These are calculated by matching acceptable acute and chronic tests and dividing the acute LC_{50} by the "Chronic Value" from the chronic test. The chronic value in turn is calculated as the geometric mean of the highest tested concentration in which selected responses were not statistically significantly different from the controls, called the no observed effect concentration (NOEC), and the lowest concentration that was statistically different from the controls, called the lowest observed effect concentration (LOEC). The selected responses considered are survival, growth, and reproduction, data on other sublethal effects such as swimming performance, or altered behaviors are put aside. The available ACRs are then selectively averaged, for a Final ACR for the substance. The continuous criterion concentration (CCC), commonly called the chronic criterion then becomes the FAV divided by the final ACR (Stephan *et al.* 1985).

This synopsis reflects the most common way the Guidelines were used with the criteria evaluated in the Opinion, but obviously doesn't reflect all the details of Stephan *et al.*'s (1985) 98 page document.

These steps and other key judgments and practices from the EPA Guidelines for developing aquatic life criteria are critically evaluated in the following parts of this section.

2.4.1.1. The assumption that not harming more than 5% of the species tested in laboratories is sufficient protection of ESA-listed species and critical habitats

The EPA's fundamental approach to setting criteria involves compiling reports of laboratory tests for species and genus mean values, rank ordering the genus mean values, and basing criteria on the 5th percentile of a distribution of the rank ordered values. This approach has been the subject of much criticism and controversy in the ecotoxicology literature. Many arguments relate to further inherent assumptions required of the approach that may not be met, are untested, or are untestable. Published concerns include:

- Whether haphazard collections of data from single-species laboratory toxicity tests can be considered relevant to natural ecosystems;

- Small datasets can be significantly biased toward more or less sensitive species than would be expected in natural ecosystems;

- Whether any species loss from a community due to a toxin is acceptable. Reducing community integrity to a simple proportion of species could discount keystone or dominant species if they were in the lower 5th percentile of sensitivity;

- Whether the 5th percentile of the SSD as the appropriate level of protection is a scientifically sound number or just a familiar number;

- Because the approach depends on comparable data, it is biased toward mortality data (which are most abundant) and biased against less abundant data on abnormal behavior or other sublethal data that may be as important for maintaining biological integrity and more relevant at low, ambient concentrations;

- The few species for which multiple tests results are available sometimes show high variability in sensitivity, yet this variability is often omitted from SSD presentations, which implies greater precision than is the case. Thus apparent differences between species' ranks on a SSD may not be meaningful, especially for species with only single or few datapoints; and

- Uncertainties in the statistical properties of the distributions and appropriate models.

(Cairns 1986; Forbes and Forbes 1993; Hopkin 1993; Smith and Cairns 1993; Underwood 1995; Power and McCarty 1997; Aldenberg and Jaworska 2000; Newman *et al*. 2000; Forbes and Calow 2002; Suter *et al*. 2002; Duboudin *et al*. 2004; Brix *et al*. 2005; Maltby *et al*. 2005; Forbes *et al*. 2008)

In contrast to these many criticisms, other studies or reviews have found reasonably good agreement between effects in laboratory and field tests (Geckler *et al*. 1976; de Vlaming and Norberg-King 1999), and lack of pronounced adverse effects in ecosystem tests at criteria-like concentrations below the 5th percentiles of SSDs (Versteeg *et al*. 1999; Mebane 2010).

No explicit consideration of protection of exceptionally vulnerable populations of threatened or endangered species was included in the criteria guidelines. However, it is clear from contemporaneous and subsequent writings by the authors that they thought criteria should specifically protect or be adjusted to protect socially valued special status species, including threatened and endangered species. For instance, the introduction to the Guidelines states that "*to be acceptable to the public and useful in field situations, protection of aquatic organisms and their uses should be defined as prevention of unacceptable long-term and short-term effects on (1) commercially, recreationally, and other important species....*" as well as fish and invertebrate assemblages (Stephan *et al.* 1985). Other writings and guidance are more explicit about the need to consider protection of species listed under the ESA; suggesting a review of whether the 95% of protected species included listed species and adequate prey for them (Stephan 1985, 1986; EPA 1994). If not, the criteria should be adjusted to protect these "critical" species. Such reviews and adjustments were recommended to be done on a site-specific basis, where a "site" may be a state, region, watershed, water body, or segment of a water body (EPA 1994). The recommendation to consider listed species at the "site" rather than national level was not stated but presumably related to complexity and the fact that imperiled species often have limited distributions.

2.4.1.2. The assumption that effects in laboratory tests are reasonable predictors of effects in field situations

The preceding discussion concerned whether compilations of laboratory test values were appropriate to treat as surrogates of the diversity of natural systems. A related but even more fundamental question is, whether tests of chemicals in laboratory aquaria with "domesticated" cultures of test animals are likely to produce similar effects as would exposure to the same substance on the same or closely related species in the wild? If the responses between animals in laboratory aquaria or the wild are different, is there likely a bias in the sensitivity of responses from either the lab or wild settings? That is, are the effects of chemical contamination more likely to be more or less severe in the laboratory or wild settings? This question is important because water quality criteria are designed to apply to and protect ambient waters, that is, streams, rivers, and lakes, yet the data used to develop them are invariably compiled from laboratory testing under tightly controlled and thus quite artificial environments.

While by definition, laboratory toxicity testing is conducted in controlled, artificial condition rather than in the wild under uncontrolled conditions, some laboratory tests are designed such that they are of questionable environmental relevance. By "environmentally relevant" in the context of interpreting laboratory toxicity tests we mean whether the test conditions were designed in a way to be relevant to conditions that might occur in the environment. Whether or not test data were environmentally relevant include the questions such as: Were fish or other organisms exposed to chemicals in concentrations ranges and ratios that actually occur in the environment? Or were organisms exposed to conditions contrived to produce effects, such as massive doses over short time periods? Were organisms exposed in a manner similar to that in the wild such as by water across the gills or diet? Or were organisms exposed in a manner designed to produce effects but wouldn't occur outside of laboratories, such as injection or a bollus in feed? In feeding studies, were chemicals in a form similar to that that might be encountered in ambient conditions? In water studies, was the dilution water a natural water type, rather than a preparation with mineral content unlike that that would occur in nature? "Environmental relevance" cannot be a hard and fast test, because studies would then be limited to field studies, which have the converse problem of being uncontrolled and difficult to unambiguously attribute apparent effects to causes. However, some studies clearly have little direct environmental relevance, and these studies are given less reliance in this opinion than "environmentally relevant" studies. For instance, in vitro tests using excised tissues, or cell lines bathed in a dosed solution are often valuable for investigations comparative biochemistry orphysiology, or on mechanisms of toxicity, but standing alone, have little direct relevance responses of a whole, living organism under conditions experienced in the wild.

There are myriad of factors that may influence the effects of a chemical stressor on aquatic organisms, and this complexity makes the question of bias in sensitivity difficult or even impossible to answer with any certainty. A number of reasons why the effects of a chemical could be more- or less-severe on listed steelhead and salmon in laboratory or in wild settings were considered and are summarized in table 2.4.1.1.

Table 2.4.1.1. Reasons why the effects of a chemical substance could be more- or less-severe on listed steelhead and salmon in laboratory or in wild settings

Factor	Are effects likely more severe in typical lab settings or in the wild?
Environmental Conditions	
Nutritional state - acute test exposures	In the wild. In acute toxicity tests with fish fry, fish are selected for uniform size, and unusually skinny fish that might be weakened from being in poor nutritional state are culled from tests. For instance, if <90% of control fish survive the 4 days starvation of an acute toxicity test, the test may be rejected from inclusion in the criteria dataset. In the wild, not all fish can be assumed to be in optimal nutritional state. While perhaps counterintuitive, starvation can protect fish against waterborne copper exposure (Kunwar *et al*. 2009). Fish are routinely starved during acute laboratory tests of the type used in criteria development.
Nutritional state – chronic test exposures	In the wild. Fish in the wild must compete for prey and if chemicals impair fish's ability to detect and capture prey because of subtle neurological impairment, this could cause feeding shifts and reduce their competitive fitness (Riddell *et al*. 2005). Fish in chronic lab tests with waterborne chemical exposures are often fed to satiation and food pellets don't actively evade capture like live prey. Perhaps these factors dampen responses in lab settings.
Temperature	In the wild. In lab test protocols, nearly optimal test temperatures are recommended, e.g., 12°C for rainbow trout, the most commonly tested salmonid. Fish may be most resistant to chemical insults when at optimal temperatures. At temperatures well above optimal ranges, increased toxicity from chemicals often results from increased metabolic rates (Sprague 1985). Under colder temperatures fish have been shown to be more susceptible to at least Cu, Zn, Se and cyanide, although the mechanisms of toxicity are unclear (Hodson and Sprague 1975; Kovacs and Leduc 1982b; Dixon and Hilton 1985; Erickson *et al*. 1987; Lemly 1993b; Hansen *et al*. 2002a).
Flow	In the wild. Fish expend energy to hold their position in streams and to compete for and defend preferred positions that provide optimal feeding opportunity from the drift for the energy expended. Subordinate fish are forced to less profitable positions and become disadvantaged. Subordinate fish in lab settings still get adequate nutrition from feeding. Chemical exposure can reduce swimming stamina or speeds, as can exposure to soft water. Chemical exposures in soft water can be expected to exacerbate effects (Adams 1975; Kovacs and Leduc 1982b; McGeer *et al*. 2000; De Boeck *et al*. 2006).
Disease and parasites	In the wild. Disease and parasite burden are common in wild fish, but toxicity tests that used diseased fish are likely to be considered compromised and results would not be used in criteria compilations. Chemical exposure may weaken immune responses and increase morbidity or deaths (Stevens 1977; Arkoosh *et al*. 1998a,b).
Predation	In the wild. Fish use chemical cues to detect and evade predators; these can be compromised by some chemical exposures (Berejikian *et al*. 1999; Phillips 2003; Scott *et al*. 2003; Labenia *et al*. 2007).

Factor	Are effects likely more severe in typical lab settings or in the wild?
Exposure	
Variable exposures	In the lab. Most toxicity tests used to develop criteria are conducted at nearly constant exposures. Criteria are expressed not just as a concentration but also with an allowed frequency and duration of allowed exceedences. In field settings, most point or non-point pollution scenarios that rarely if ever exceed the criteria concentration (i.e., no more than for one four day interval per 3 years), will have an average concentration that is less than the criteria concentration. For some chemicals, such as copper, fish might detect and avoid harmful concentrations if clean-water refugia were readily available.
Metal form and bioavailability	Uncertain. Metals other than Hg and some organics are commonly assumed to be more bioavailable in the lab because dissolved organic carbon (DOC), which reduces the bioavailability and toxicity of several metals, is low in laboratory tests that are eligible for use in criteria. The Guidelines call for <5 mg/L TOC (total organic carbon) in order to be used in criteria (Stephan *et al*. 1985), but probably more often TOC is <2 mg/L in laboratory studies However, in mountainous streams in Idaho, TOC is often as low (≈1-2 mg/L) during baseflow conditions (Appendix C), so differences in bioavailability between streams and laboratory waters that both have low TOC are not necessarily large. (Organic carbon is more often discussed as DOC in this Opinion. TOC includes particulates, which other than during runoff conditions in streams will tend to be low and thus TOC and DOC would be similar during conditions without runoff).
Chemical equilibria	Uncertain. While results conflict, metals are usually considered less toxic when in equilibrium with other constituents in water, such as organic carbon, calcium, carbonates and other minerals. In the wild, daily pH cycles prevent full equilibria from being reached (Meyer *et al*. 2007a). Likewise, in conventional laboratory flow-through test designs chemicals may not have long enough contact time to reach equilibria. Static-renewal tests are probably nearly in chemical equilibria although organic carbon accretion can lessen toxicity which may not reflect natural settings (Santore *et al*. 2001; Welsh *et al*. 2008).
Prior exposure	Uncertain. If fish are exposed to sublethal concentration of a chemical, they could potentially either become weakened or become more tolerant of future exposures. With some metals, normally sensitive life stages of fish may become acclimated and less sensitive during the course of a chronic test if the exposure was started during the resistant egg stage (Chapman 1983, 1985; Sprague 1985; Brinkman and Hansen 2007). (further discussion follows in the text).
Life stages exposed	In the wild. Most lab studies are short term; realistically testing all life stages of anadromous fish is probably infeasible. Reproduction is often the most sensitive life stage with fish but most "chronic" studies are much shorter and just test early life stage survival and growth (Suter *et al*. 1987). At different life stages and sizes, salmonids can have very different susceptibility to some chemicals; even when limited to a narrow window of YOY fry, sensitivity can vary substantially (this review). Unless the most sensitive life stages are tested, lab tests could provide misleadingly high toxicity values for listed species (further discussion follows in the text).

Factor	Are effects likely more severe in typical lab settings or in the wild?
Chemical mixtures	In the wild. In field conditions, organisms never experience exposure to a single pollutant; rather, ambient waters typically have low concentrations of numerous chemicals. The toxic effects of chemicals in mixture can be less than those of the same chemicals singly, greater than, or have no appreciable difference. The best known case of one toxicant reducing the effects of another is probably Se and Hg (e.g., Belzile *et al.* 2006). However, strongly antagonistic responses are probably uncommon, and much more common are situations where chemical mixtures have greater toxicity than each singly or little obvious interaction (e.g., Norwood *et al.* 2003; Borgert 2004; Playle 2004; Scholz *et al.* 2006; Laetz *et al.* 2009). In general, it seems prudent to assume that if more than one toxicant were jointly elevated it is likely that lower concentrations of chemicals would be required to produce a given magnitude of effect than would be predicted from their actions separately. However, the magnitude or increased effects at environmentally relevant concentrations is uncertain and for some combinations may be slight or imperceptible.
Dietary exposures	In the wild. Toxicity test data used in criteria development have been mostly based solely on waterborne exposures, yet in the wild, organisms would be exposed to contaminants both through dietary and water exposures. With at least some organics (e.g., dioxins, PCBs) dietary exposures are more important than water exposures as is the case for some inorganics (As, Hg, Se). For some other metals (Cd, Cu, Ni, Pb, Zn), at environmentally relevant concentrations that would be expected when waterborne concentrations are close to criteria, dietary exposures have not been shown to directly result in appreciable adverse effects to fish (Hansen *et al.* 2004; Schlekat *et al.* 2005; Erickson *et al.* 2010). However, while dietary exposures of metals have not yet been implicated in adverse effect to fish at or below criteria concentrations, they may in fact be both the primary route of exposure and an important source of toxicity for benthic invertebrates (Irving *et al.* 2003; Poteat and Buchwalter 2014). For instance Besser *et al.* (2005a) found that the effects threshold for Pb to the benthic crustacean *Hyalella* was well above the chronic criterion in water exposures, but when Pb was added to the diet, effects threshold dropped to near criteria concentrations. Ball *et al.* (2006) found that feeding Cd contaminated green algae to the benthic crustacean *Hyalella* caused a 50% growth reduction at about the NTR chronic criteria.
Population dynamics	
Density effects	In the lab. Salmonid fishes are highly fecund (~500 to 5000 eggs per spawning female). When abundant, overcrowding and competition for food and shelter may result in relatively high death rates for some life stages, particularly YOY during their first winter. After many fish die in a density-dependent bottleneck, the survivors have greater resources and improved growth and survival. Conceptually, if an acute contamination episode killed off a significant portion of YOY fish prior to their entering a resource bottleneck, then assuming no residual contaminant effects, the losses to later life stages and to adult spawners would be buffered.
Meta-population dynamics	In the lab. If habitats are interconnected, as is the case in intact stream networks, then if pervasive contamination from discharges to a stream were to impair only some endpoints or life-stages, such as reproductive failure or YOY mortalities, immigration from source populations may make detection of population reductions in the affected sink population difficult (Ball *et al.* 2006; Palace *et al.* 2007). If an episodic contamination pulse were to kill a large proportion of fish in a stream, the proximity of refugia and donors from source populations affect recovery rates (Detenbeck *et al.* 1992).

Considering all the reasons why the effects of a given chemical concentration could have more or less severe effects in laboratory settings or the wild, general conclusions are elusive. It may be that the best overall conclusion is the same as that reached by Chapman (1983) that "*when appropriate test parameters are chosen, the response of laboratory organisms is a reasonable index of the response of naturally occurring organisms.*" His conclusion in turn contributed to one the most fundamental assumptions of EPA Guidelines, that is, "*these National Guidelines have been developed on the theory that effects which occur on a species in appropriate laboratory tests will generally occur on the same species in comparable field situations.*"

Summary: Based on this analysis, the assumption that effects in laboratory tests are reasonable predictors of effects to species in the wild is dependent upon the specific factor being considered. While it is generally reasonable to interpret effects from laboratory tests as being applicable to field situations where criteria are applied, there is some risk that laboratory tests may underpredict effects in the wild.

2.4.1.3. Susceptibility of Salmonids to Chemicals at Different Life Stages

Since a species can only be considered protected from acute toxicity if all life stages are protected, EPA's Guidelines recommend that if the available data indicate that some life stages are at least a factor of two more resistant than other life stages, the data for the more resistant life stages should not be used to calculate species mean acute values (Stephan *et al*. 1985). Smaller, juvenile life stages of fish are commonly expected to be more vulnerable to metals toxicity than larger, older life stages of the same species. For instance, a standard guide for testing the acute toxicity of fish recommends that tests should be conducted with juvenile fish, that is, post-larval or older and actively feeding, usually in the size range from 0.1 and 5.0g in weight (ASTM 1997).

A review of several data sets in which salmonids of different sizes were similarly tested shows that even among juvenile fish in the 0.1 to 5.0g size range, differences in sensitivity can approach a factor of 10. This emphasizes the importance of EPA's guidance not to use the more resistant life stages. However, the data sets analyzed indicated that in practice, there were sometimes greater influences of life stage on the sensitivity of salmonids to some substances than was apparent to the authors of the individual criteria documents using the datasets available to them at the time. Some of the SMAVs and GMAVs which were used to rank species sensitivity and set criteria were considerably higher than EC_{50}s with salmonids that were tested at the most sensitive life stages (Figures 2.4.1.1 to 2.4.1.4).

For three Pacific salmonid species for which comparable test data were available for different life stages; coho salmon (*Oncorhynchus kisutch*), rainbow trout (*O. mykiss*) and cutthroat trout (*O. clarki*), the data suggest that swim-up fish weighing around 0.5g to about 1g may be the most sensitive life stage. None of the data sets examined in detail or other published studies reviewed had sufficient resolution to truly define at what weight fish became most sensitive to metals, but along with other data they suggest that larger fish may be less sensitive than fish at 0.4 to 0.5g. For instance with zinc, rainbow trout in the size range of about 0.1 to about 1.5 g consistently became more sensitive to zinc in two studies with multiple tests in that size range (Figure 2.4.1.2

and Figure 2.4.1.3). The paucity of data with salmonids in the size range of about 0.5 to 2g prevents definitive statements of a most sensitive size across species or even tests. All data located for early swim-up stage *Oncorhynchus* in the 0.1 to 0.5g range were consistent with increasing sensitivity with size. With Hansen *et al's*. (2002c) rainbow trout studies, this relationship continued with fish up to about 1.5g. However, with cutthroat trout, the few data available suggests that fish larger than about 0.5g become less sensitive with increasing size (Figure 2.1.4.2).

Some studies with older and larger rainbow trout have found that the fish became more resistant to zinc and copper (Chapman 1978b; Chapman and Stevens 1978; Howarth and Sprague 1978; Chakoumakos *et al*. 1979). Studies with copper all showed this trend, but the strength of size-sensitivity relations varied across studies. Chakoumakos *et al*. (1979) found that fish between about 1 and 25g in weight varied in their sensitivity to copper by about eight times (Figure 2.4.1.4), but steelhead (*O. mykiss*) that were tested with copper at sizes of 0.2, 7, 70, and 2700g showed little pattern of sensitivity with size (Chapman 1978b; Chapman and Stevens 1978). However, the large differences in sizes may have missed changes at intermediate sizes in the ranges compared at Figures 2.4.1.1 to 2.4.1.4. Similarly, with copper and rainbow trout, Anderson and Spear (1980) found that three sizes of rainbow trout (3.9, 29 and 176g) had similar sensitivities.

NMFS reviewed several data sets that indicated increasing susceptibility of salmonids to at least metals with increasing size and age as fish progressed from the resistant alevin stage. The "U" shaped size-sensitivity response with the most sensitive life stage for salmonids fish around 0.5g in weight seems a reasonable interpretation of the available data, but few data were available in the size range of 0.5 to 2g, so it is possible the most sensitive stage is larger. Hedtke *et al*. (1982) tested coho salmon for the influence of body size and developmental stage with copper, zinc, nickel, and PCP. Fish were exposed as alevins, swim-up fry, and juveniles, and within these developmental stages smaller fish were tested against larger fish. For copper, zinc, and PCP, the swim-up fry stage was most susceptible, and within the swim-up stage, the larger fish were more susceptible to copper and zinc than smaller fish (~0.25g vs. 0.7g fish, wet weight). For PCP, there was no difference for size of fish within the sensitive alevin to swim-up stage, and with Ni all fish were very resistant (Hedtke *et al*. 1982). In three test pairs with rainbow trout exposed to cadmium and zinc under similar hardness, pH, and temperature, the fish tended to become more sensitive with increasing size from 0.4 to 0.9g for rainbow trout and zinc, and 0.26 to 0.66g with Cd. Further growth in juvenile rainbow up to 1.1 and 1.6g for cadmium and zinc had little effect on sensitivity (Figure 2.4.1.3). In parallel tests with bull trout (*Salvelinus confluentus)*, size had little effect on sensitivity over a range of 0.08 to 0.22g for cadmium although with zinc; however, the smallest fish (0.1g) were also least sensitive (Hansen *et al*. 2002c). Similar tests with copper and rainbow and bull trout showed roughly similar patterns. Three tests with rainbow trout at the same hardness and using fish from the same source had the most sensitive results for 0.43g fish (LC_{50}s of 36, 54, and 93 µg/L for rainbows weighing 0.43, 0.3, and 0.68g, respectively). Bull trout tested at constant temperature of 8°C tended to become more sensitive with increasing size up to ~1g (Hansen *et al*. 2002a). Besser *et al*. (2007) similarly found that 0.5g rainbow trout were more sensitive than 0.13g fish to copper and zinc, but not for cadmium.

These patterns do not seem to hold for all species. Contrary to the patterns with the salmonids, newly hatched sculpins were more sensitive to cadmium, copper, and zinc than were older juveniles (Besser *et al*. 2007). Similar to the sculpin results but contrary to all the other salmonid results, Carney *et al*. (2008) found that the brown trout (*Salmo trutta*) became less sensitive to copper with increasing size. Guppies exposed to toxicants with different modes of action tended to become more susceptible with increasing size and age (dieldrin, PCP, cyanide, copper, zinc, and nickel) (Anderson and Weber 1975).

Summary: Salmonids can have profound differences in susceptibility to chemicals at different life stages, and in some instances, species mean acute values used in criteria may be skewed high because insensitive life stages were included. A "U" shaped pattern of sensitivity with life stage was suggested for several datasets with Pacific salmon or trout species (i.e., *Oncorhynchus)* and some metals. Across several good datasets, the most vulnerable life stage and size appeared to be swim-up fry weighing between about 0.5 to 1.5g. However, no consistent pattern was obvious across other species of fish, chemicals, and life stages.

Caution is needed when using SMAVs or GMAVs as summary statistics for ranking species sensitivity or setting criteria. Reviews of the protectiveness of chemical concentrations or criteria that rely in large part upon published mean acute values for species of special concern such threatened species, or their surrogates, may be subject to considerable error if the underlying data points are not examined. This may include analyses such as SSD, interspecies correlation estimates (ICE, Asfaw *et al*. (2004), or any other relative sensitivity comparisons that uses mean acute values at the family, genus, or species level.

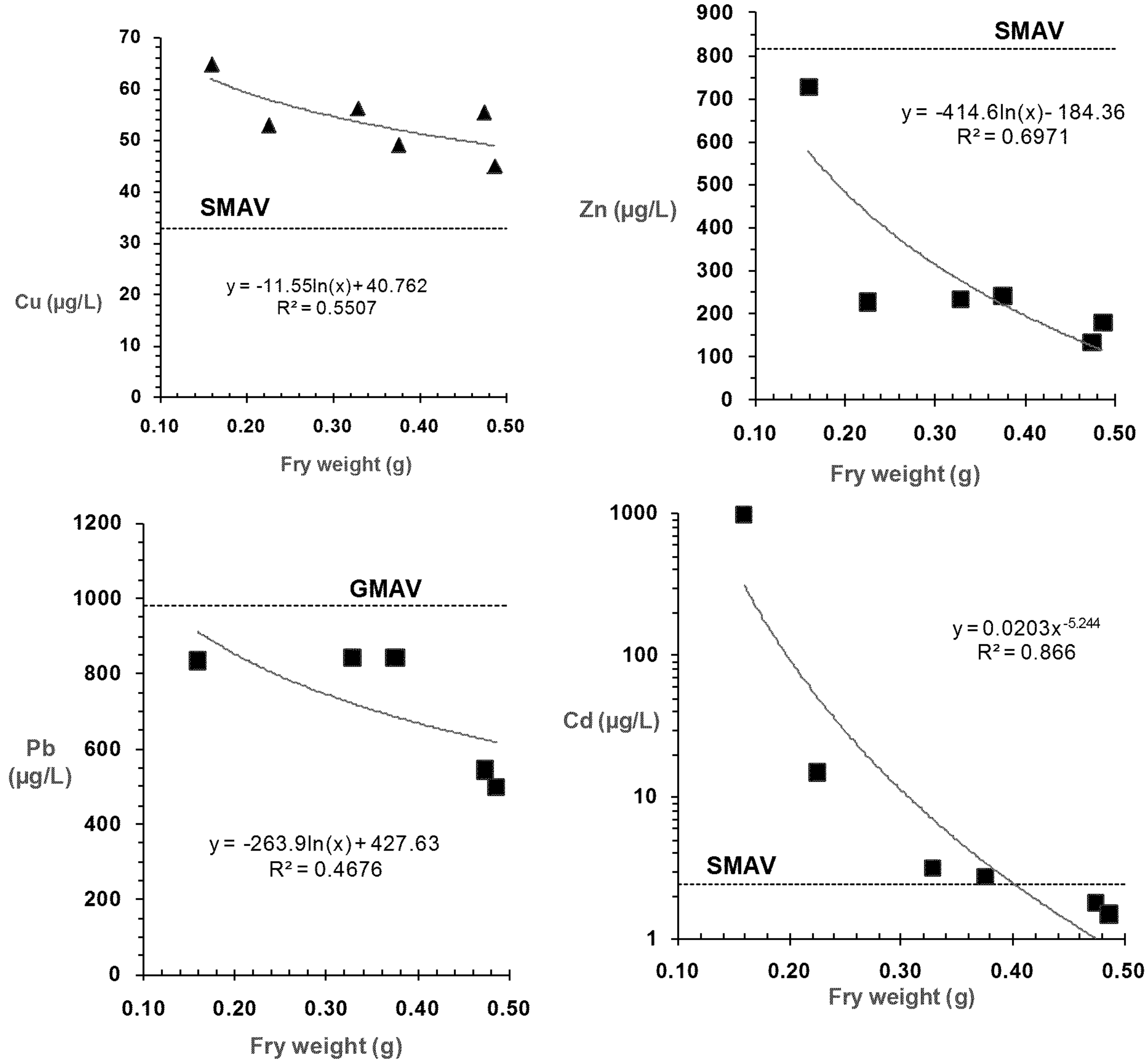


Figure 2.4.1.1. Size-developmental stage patterns with coho salmon from 2 to 7 weeks post hatch, data from Chapman (1975). Species and genus mean acute values (SMAVs and GMAV) are from the respective criteria documents (EPA 1984b, 1984a, 1985, 1987b), adjusted to test water hardness. All tests used Willamette River water, TOC 3.4 mg/L, hardness 22 mg/L.

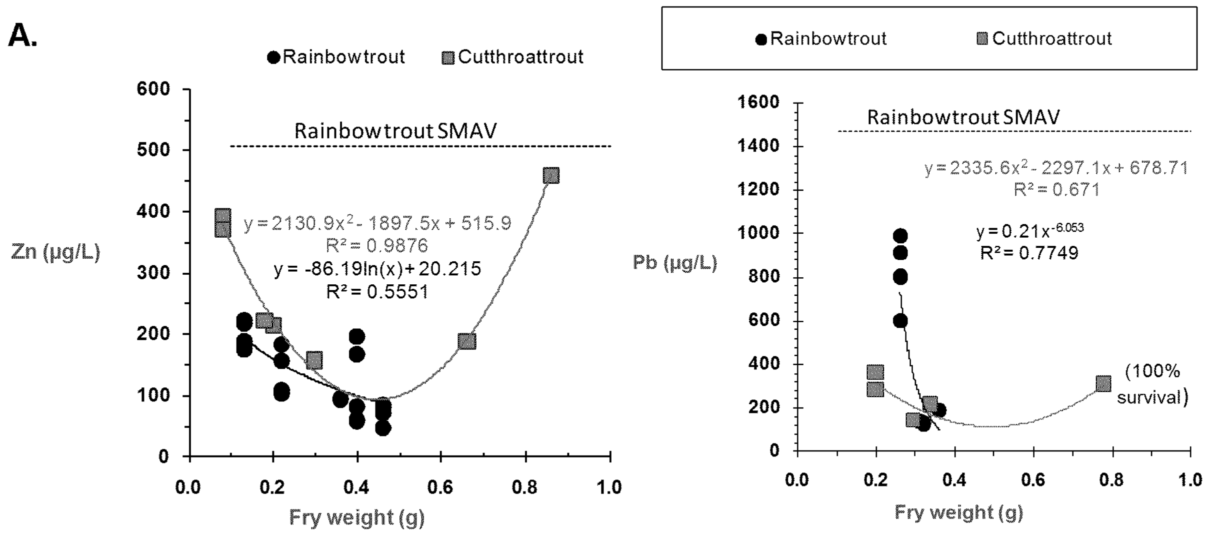


Figure 2.4.1.2. Relations between size of swim-up rainbow and cutthroat trout and toxicity to zinc and lead sensitivity in renewal tests conducted in water from the South Fork Coeur d'Alene River, Idaho. Data from (Mebane *et al.* 2012). All test values adjusted to a median test hardness of 35 mg/L $CaCO_3$ using hardness-toxicity regressions from (Mebane *et al.* 2012). SMAVs were adjusted using the hardness-criteria equations from the respective criteria documents.

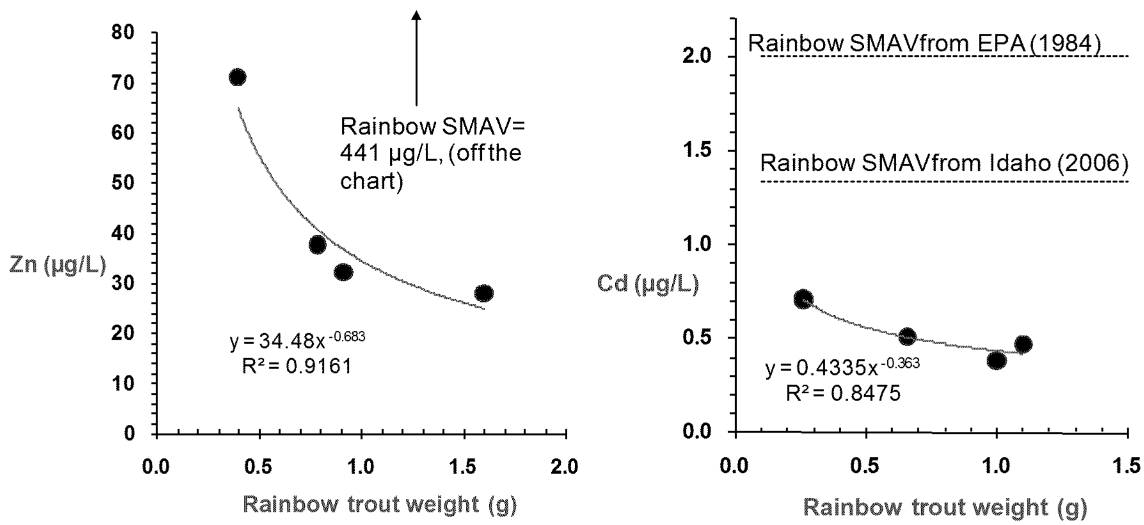


Figure 2.4.1.3. Resistance to cadmium and zinc toxicity decreased with increasing size over a weight range of 0.2 to 1.6g for swim-up rainbow trout. Data from Hansen (2002a) and Stratus (1999) using 96-h probit LC_{50} values. All tests conducted at a hardness of 30 mg/L and pH of 7.5 SMAV values were adjusted using the hardness-criteria equations from the respective criteria documents.

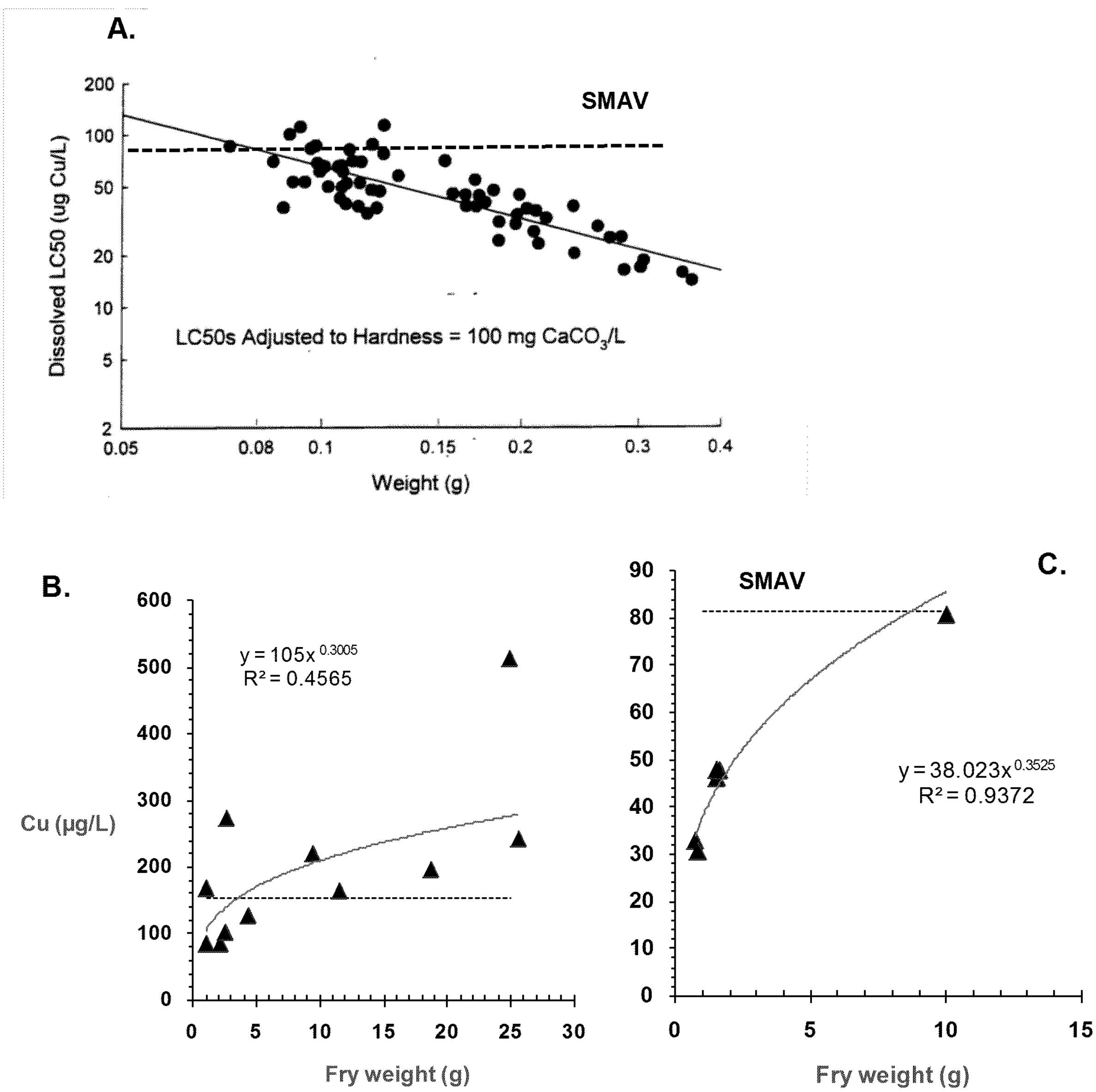


Figure 2.4.1.4. Resistance to copper toxicity decreased with increasing size over a weight range of 0.06 to 0.4g for swim-up rainbow trout, but above about 1g weight, resistance to copper toxicity increased with increasing size. Dashed lines indicate hardness-adjusted rainbow trout species mean acute value (SMAV) from EPA (1984). A. Relation between copper toxicity and the size of swim-up rainbow trout (<0.5g), from renewal tests conducted in water from the Clark Fork River, MT (Erickson *et al.* 1999); B. Relation between copper toxicity and the size of larger juvenile rainbow trout (>0.7g, older than swim-up fish), data from Chakoumakos *et al*'s (1979) tests under uniform water conditions (hardness 194 mg/L); C. Rainbow trout of difference sizes tested under uniform conditions at hardness 99 to 102 mg/L, data from Howarth and Sprague (1978).

2.4.1.4. Effects of Acclimation on Susceptibility to Chemicals

Exposure to sublethal concentrations of organic chemicals and other metals may result in pronounced increases in resistance to later exposures of the organisms. With metals, the increased resistance may be on the order of two to four times for acute exposures, but may be much higher for some organic contaminants (Chapman 1985). However, the increased resistance can be temporary and can be lost in as little as 7 days after return to unpolluted waters (Bradley *et al*. 1985; Sprague 1985; Hollis *et al*. 1999; Stubblefield *et al*. 1999). For this reason, EPA's Guidelines specify that test results from organisms that were pre-exposed to toxicants should not be used in criteria derivation (Stephan *et al*. 1985).

However, there is a less obvious source of acclimation that is not precluded by the Guidelines and influences chronic values and thus chronic criteria. Several tests have shown that life stages typically sensitive to toxins (e.g., fry stage) become more resistant when toxicity tests were initiated during resistant early life stages (ELS, e.g., embryo stage). This suggests that acclimation to toxin(s) during ELS exposure may lead to greater resistance in later life stages in comparison to the same life stages of naïve fish (fish which had no previous exposure) (Chapman 1978a; Spehar *et al*. 1978; Chapman 1994; Brinkman and Hansen 2004, 2007). The Guidelines could actually be interpreted to exclude chronic exposures that did not pre-expose, and acclimate fish to metals as eggs (Stephan *et al*. 1985), which was probably unintended.

Chapman (1994) exposed different life stages of steelhead (*Oncorhynchus mykiss*) for the same duration (3 months) to the same concentration of copper (13.4 µg/L at a hardness of 24 mg/L as $CaCO_3$). The survival of steelhead which were initially exposed as embryos was no different from that of the unexposed control fish, even though the embryos developed into the usually-sensitive swim-up fry stage during the exposure. In contrast, steelhead which were initially exposed as swim-up fry without the opportunity for acclimation during the embryo state, suffered complete mortality *(*Figure 2.4.1.4*)*. Brinkman and Hansen (2007) compared the responses of brown trout (*Salmo trutta*) to long-term cadmium exposures that were initiated either at the embryo stage (i.e., ELS tests) or the swim-up fry stage (i.e., chronic growth and survival tests). In three comparative tests, fish that were initially exposed at the swim-up fry stage were consistently two to three times less resistant than were the fish initially exposed at the embryo stage.

These studies support the counterintuitive conclusion that because of acclimation, longer-term tests or tests that expose fish over their full life cycle are not necessarily more sensitive than shorter-term tests which are initiated at the sensitive fry stage. Conceptually, whether this phenomenon is important depends on the assumed exposure scenario. If it were assumed that spawning habitats would be exposed, then the less-sensitive ELS tests would be relevant. However, for migratory fishes such as listed salmon and steelhead, their life histories often involve spawning migrations to headwater reaches of streams, followed downstream movements of fry shortly after emerging from the substrates, and followed by further seasonal movements to larger, downstream waters to overwinter (Willson 1997; Baxter 2002; Quinn 2005). These life history patterns often correspond to human development and metals pollution patterns such that headwater reaches likely have the lowest metals concentrations, and downstream increases could occur due to point source discharges or urbanization.

From the discussion in the Guidelines of the types of chronic data with fish that are acceptable for use in criteria development, it is clear that the intent was to capture information on the most sensitive life stage of a fish species. Unfortunately, the wording of the Guidelines could be interpreted to preclude the use of the more sensitive chronic growth and survival tests that were initiated with salmonid fry stage, and specify the use of the less sensitive ELS tests (Stephan *et al*. 1985, at p. 44).

Summary: In chronic tests with salmonids and metals, the Guidelines inadvertently favor a test method (ELS tests) that may be inherently biased toward insensitivity because acclimation can occur during the insensitive egg stage of exposure. Thus, Species Mean Chronic Values listed in criteria documents may be also be biased high.

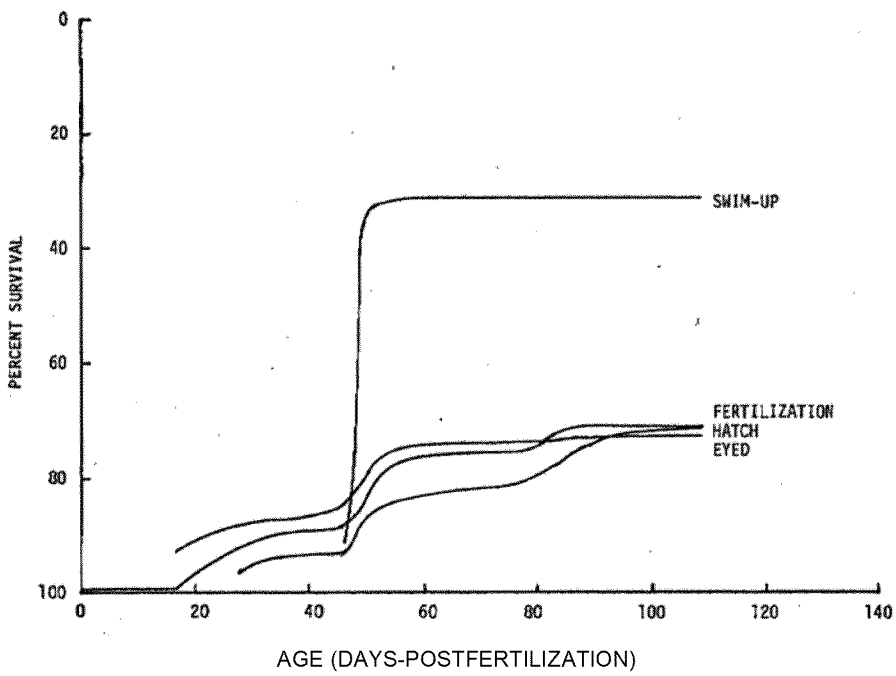


Figure 2.4.1.5. Effect of developmental stage at the onset of continuous copper exposure (13.4 µg/L) on the survival of juvenile steelhead trout (figure from Chapman 1994).

2.4.1.5. Implications of the use of the "chronic value" statistic in setting criteria

A related issue with the derivation of chronic criteria is the test statistic used to summarize chronic test data for species and genus sensitivity rankings. Literature on chronic effects of chemicals often contains variety of measurement endpoints, different terms, and judgments by the authors of what constitutes an acceptable or negligible effect. While the Guidelines give a great deal of advice on considerations for evaluating chronic or sublethal data (Stephan *et al.* 1985, at p.39), those considerations were not usually reflected in the individual criteria documents reviewed for this consultation. In practice for most of the criteria documents reviewed, "chronic values" were simply calculated as the geometric mean of the lowest tested concentration that had a statistically significant adverse effect at the 95% confidence level (lowest observed effects concentration [LOEC]) and the next lower tested concentration (no observed effects concentration [NOEC]). The "chronic value" as used in individual criteria documents is effectively the same thing as the maximum acceptable toxicant concentration (MATC) used in much environmental toxicology literature, even though the MATC term is never used in the Guidelines. This MATC approach has the potential to seriously underestimate effects because the statistical power in typical toxicity tests is fairly low. A bias in many ecotoxicology papers is to focus on avoiding "false accusations" of a chemical with 95% accuracy (i.e., Type I error or false positive, the risk of declaring an effect was present when in fact the apparent effects only occurred by chance). Often no consideration whatsoever is given to the companion problem, known as Type II error, or false negatives, (i.e., declaring no adverse effects occurred when in fact they did but because of the limited sample size or variability, were not significant with 95% confidence).

The magnitude of effect that can go undetected with 95% confidence in a NOEC statistic can be large, greater than 30% on average for some endpoints, and much higher for individual tests (Crane and Newman 2000). This problem is compounded with the "chronic value" or MATC when calculated in its most common form as the geometric mean of a NOEC and LOEC. For instance, 100% of juvenile brook died after being exposed to 17 µg/L copper for 8 months; this was considered the LOEC for the test. The next lowest concentration tested (9.5 µg/L) had no reduced survival relative to controls (McKim and Benoit 1971). Therefore, the only thing that can be said about the geometric mean of these two effect concentrations, i.e., the chronic value of 12.8 µg/L that was used in the chronic copper criteria (EPA 1985d) is that it represents a concentration that can be expected to kill somewhere between all or no brook trout in the test population. Similarly, Grosell *et al.* (2006a) showed that the NOECs and LOECs for reduced growth in snails exposed to lead corresponded with about a 57% and 90% growth reduction, and over 70% reduced growth for the MATC. Animals suffering such severe stunted growth may not even reproduce, so the MATC would not seem to be a very acceptable maximum toxicant concentration. Suter *et al.* (1987) evaluated published chronic tests with fish for a variety of chemicals and found that on the average the MATC represented about a 20% death rate and a 40% reduction in fecundity. They noted that "*although the MATC is often considered to be the threshold for effects on fish populations, it does not constitute a threshold or even a negligible level of effect in most of the published chronic tests. It corresponds to a highly variable level of effect that can only be said to fall between 0% and 90%.*" Barnthouse *et al.* (1989) further extrapolated MATC-level effects to population-level effects using fisheries sustainability models and found that the MATC systematically undervalued test responses such as fecundity, which are both highly sensitive and highly variable.

One implication of this issue is that because the MATC chronic values typically used in criteria documents under review may represent substantial adverse effects for that test species, the criteria on the whole will be less protective than the intended goal of protecting 95% of the species. How much less protective is unclear and probably varies among the criteria datasets. One dataset from which a hypothetical NOEC-based chronic criterion could readily be recalculated and compared with the usual MATC criteria was a 2006 cadmium criteria update

(Mebane 2006). In this comparison, the MATC-based chronic criteria would protect about 92% of the aquatic species in the dataset at the NOEC level. Because the NOEC statistic also can reflect a fairly sizable effect (Crane and Newman 2000), it may be that at least with Cd, the true level of protection is closer to about 90% than the 95% intended by the Guidelines.

A specific question for interpreting ecotoxicological data to evaluate the protectiveness of species listed under the ESA is, what level of effect is "insignificant?" "Insignificant effects" have been defined in this context to "*relate to the size of the impact and should never reach the scale where take occurs*" and "*based on best judgment, a person would not be able to meaningfully measure, detect, or evaluate insignificant effects*" (USFWS and NMFS 1998). To evaluate what test statistic best approximated a "true" no-effect concentration for evaluating risks to ESA-listed species, we made a limited comparison of NOECs versus regression or distribution-based methods for estimating no- or very low effects concentrations. The alternative statistics evaluated were the lower 95th percentile confidence limit of the concentration affecting 10% of the test population (LCL- EC10), or estimates of the EC1 or EC0 (1% or 0% effects). NMFS concluded that the EC0 was the preferred, best estimate of no-effect value from a toxicity test. However, if data were insufficient to calculate an EC0 or other regression based approaches, the NOEC may be the best available statistic for estimating "insignificant" effects (Appendix B).

<u>Summary:</u> The Chronic Value statistic is calculated by splitting the difference between an adverse effects concentration (the LOEC) and a concentration expected to have low adverse effects (the NOEC). However, in practice the NOEC can have more adverse effects than implied by the term "NOEC", and splitting the difference between two adverse effects concentrations produces another adverse effect concentration. Thus the Chronic Value statistic used to set chronic criteria through ACRs, etc., in practice produces an uncertain level of effect and may result in less protection than intended by the EPA Guidelines. This has been estimated to result in a level of protection was closer to about 90% of the species represented in an SSD than the 95% intended by the Guidelines.

2.4.1.6. The assumption that dividing a concentration that killed 50% of a test population by two will result in a safe concentration

One challenge for deriving aquatic life criteria for short-term (acute) exposures is that the great majority of available data is for mortality, which is a concentration that kills 50% of a test population. A fundamental assumption of EPA's criteria derivation methodology is that the FAV, the LC_{50} for a hypothetical species with a sensitivity equal to the 5th percentile of the SSD, may be divided by two in order to extrapolate from a concentration that would likely be extremely harmful to sensitive species in short-term exposures (kill 50% of the population) to a concentration expected to kill few, if any, individuals. This assumption, which must be met for acute criteria to be protective of sensitive species, is difficult to evaluate from published literature because so few studies report the data behind an LC_{50} test statistic. While LC_{50}s are almost universally used in reporting short-term toxicity testing, they are not something that can be "measured" but are statistical model fits. An acute toxicity test is actually usually a series of four to six tests run in parallel in order to test effects at different chemical concentrations. An

LC_{50} is estimated by a statistical distribution or regression model which generates an LC_{50} estimate, usually a confidence interval, and then all other information is thrown away. Thus, while the original test data included valuable information on what concentrations resulted in no, low, or severe effects, that information is lost to reviewers unless the unpublished raw lab data are available to them.

The assumption that dividing an LC_{50} by two will result in a no- or very low effects concentration rests on further assumptions of the steepness of the concentration-response slope. Several examples of tests with metals which had a range of response slopes are shown in Figure 2.4.1.6. We selected these examples from data sets that were relevant to salmonid species in Idaho and for which the necessary data to evaluate the range of responses could be located (Chapman 1975, 1978b; Marr *et al.* 1995b; Marr *et al.* 1999; Mebane *et al.* 2010; Mebane *et al.* 2012).

The citations are to reports with detailed enough original data to examine the mortality at the LC_{50} concentration divided by two. The vast majority of published data was inadequate for this comparison, because usually only the LC_{50}s are reported, not the actual responses by concentration. We examined around 100 tests for this comparison. The examples shown in Figure 2.4.1.6 range from tests with some of the shallowest concentration-response slopes located to very steep response slopes. In the shallowest tests (panels *A and E*), an $LC_{50}/2$ concentration would still result in 15% to 20% mortality. However, a more common pattern with the metals data was that an $LC_{50}/2$ concentration would probably result in about a 5% death rate (panels B and F), and in many instances, no deaths at all would be expected (panels C and D).

In one of the few additional published sources that gave relevant information, Spehar and Fiandt (1986) included effect-by-concentration information on the acute toxicity of chemical mixtures. Rainbow trout and *Ceriodaphnia dubia* were exposed for 96 and 48 hours, respectively, to a mixture of six metals, each at their presumptively "safe" acute CMC. In combination, the CMC concentrations killed 100% of rainbow trout and *Ceriodaphnia*, but 50% of the CMC concentrations killed none (Spehar and Fiandt 1986). This gives support to the assumption that dividing a lethal concentration by two would usually kill few if fish, although it does not bode well for arguments of the overall protectiveness of criteria concentrations in mixtures.

Other reviews include Dwyer *et al.* (2005b) who evaluated the "$LC_{50}/2$" assumption with the results of the acute toxicity testing of 20 species with five chemicals representing a broad range of toxic modes of action. In those data, multiplying the LC_{50} by a factor of 0.56 resulted in a low (10%) or no-acute effect concentration. Testing with cutthroat trout and cadmium, lead, and zinc singly and in mixtures, Dillon and Mebane (2002) found that the $LC_{50}/2$ concentration corresponded with death rates of 0% to 15%.

<u>Summary:</u> The assumption that one-half of an LC_{50} concentration for a sensitive test, i.e., a concentration near the 5^{th} percentile of the ranked species sensitivities, will result in little or no deaths was supported by several data sets plus two published articles. While up to 20% mortality was calculated, in most cases the expected morality associated with a $LC_{50}/2$ was less than 10% and often zero.

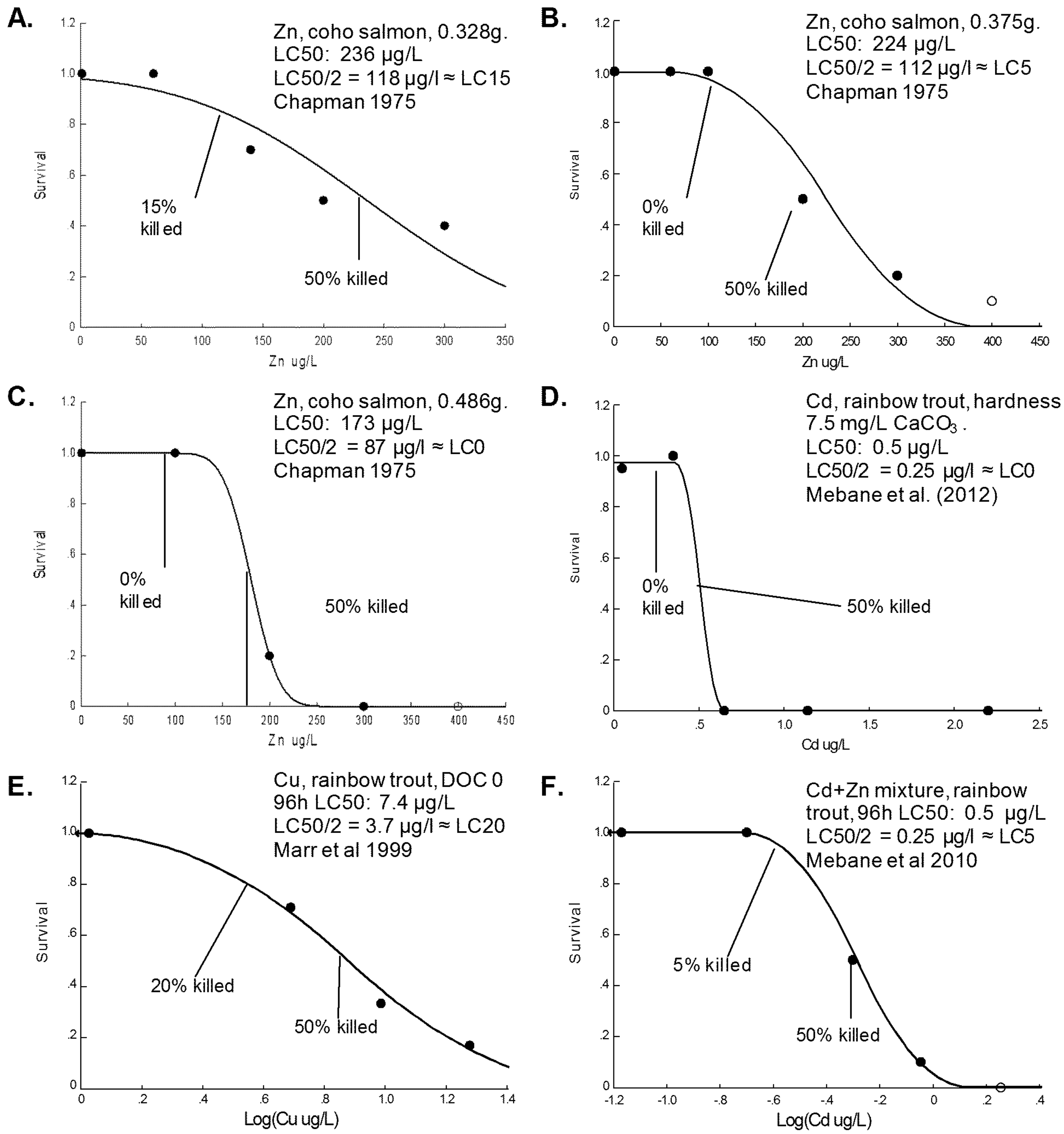


Figure 2.4.1.6. Examples of percentages of coho salmon or rainbow trout killed at one-half their LC_{50} concentrations with cadmium, copper, and zinc.

2.4.1.7. Issue of Using Flow Through, Renewal, or Static Exposure Test Designs

One area of controversy in evaluating toxicity test data or risk assessments or criteria derived from them has to do with potential bias in how test organisms are exposed to test solutions. Exposures of test organisms to test solutions are usually conducted by variations on three techniques. In "static" exposures test, solutions and organisms are placed in chambers and kept there for the duration of the test. The "renewal" technique is like the static technique except that test organisms are periodically exposed to fresh test solution of the same composition, usually once every 24 or 48 hours, by replacing nearly all the test solution. In the "flow-through" technique, test solution flows through the test chamber on a once-through basis throughout the test, usually with at least five volume replacements/day (ASTM 1997).

The term "flow-through test" is commonly mistaken for a test with flowing water, i.e., to mimic a lotic environment in an artificial stream channel or flume. This is not the case; rather the term refers to the once-through, continuous delivery of test solutions (or frequent delivery in designs using a metering system that cycles every few minutes). Flows on the order of about 5-volume replacements per 24 hours are insufficient to cause discernible flow velocities. In contrast, even very slow moving streams have velocities of around 0.04 ft/sec (a half inch per second) or more. At that rate, a parcel of water would pass the length of a standard test aquarium (~2 ft) in about 48 seconds, resulting in about 3,600 volume replacements per day. At more typical stream velocities of about 0.5 ft/sec would produce over 20,000 volume replacements/day.

Historically, flow-through toxicity tests were believed to provide a better estimate of toxicity than static or renewal toxicity tests because they provide a greater control of toxicant concentrations, minimize changes in water quality, and reduce accumulation of waste products in test exposure waters (Rand *et al*. 1995). Flow-through exposures have been preferred in the development of standard testing protocols and water quality criteria. The EPA Guidelines first advise that for some highly volatile, hydrolysable, or degradable materials, it is probably appropriate to use only results of flow-through tests. However, this advice is followed by specific instructions that if toxicity test results for a species were available from both flow-through and renewal or static methods, then results from renewal or static tests are to be discounted (Stephan *et al*. 1985). Thus, depending upon data availability, toxicity results in the criteria databases may be a mixture of data from flow through, renewal, or static tests, raising the question of whether this could result in bias. In the 1985 Guidelines, the rationale for the general preference for flow-through exposures was not detailed, but it was probably based upon assumptions that static exposures will result in LC_{50}s that are biased high (apparently less toxic) than comparable flow-through tests or because flow-through tests are assumed have more stable exposure chemistries and will result in more precise LC_{50} estimates.

With metals, renewal tests have been shown to produce higher EC_{50}s (i.e., metals were less toxic), probably because of accretion of dissolved organic carbon (DOC) (Erickson *et al*. 1996; Erickson *et al*. 1998; Welsh *et al*. 2008). However, in contrast to earlier EPA and American Society for Testing and Materials (ASTM) recommendations favoring flow-through testing, Santore and others (2001) suggested that flow-through tests were biased low because copper complexation with organic carbon, which reduces acute toxicity, is not instantaneous and typical flow-through exposure systems allowed insufficient hydraulic residence time for complete copper-organic carbon complexation to occur. Davies and Brinkman (1994) similarly found that cadmium and carbonate complexation was incomplete in typical flow-through designs, although in their study incomplete complexation had the opposite effect of the copper studies, with cadmium in the aged, equilibrium waters being more toxic. A further complication is that it is not at all clear that natural flowing waters should be assumed to be in chemical equilibria because of tributary inputs; hyporheic exchanges; and daily pH, inorganic carbon, and temperature cycles. Predicting or even evaluating risk of toxicity through these cycles is complex and seldom attempted (Meyer *et al.* 2007a), in part because pulse exposures cause latent mortality (i.e., fish die after exposure to the contaminant is removed), a phenomenon that is often overlooked or not even recognized in standard acute toxicity testing.

When comparing data across different tests, it appears that other factors such as testing the most sensitive sized organisms or organism loading may be much more important than if the test was conducted by flow through or renewal techniques. For instance, Pickering and Gast's (1972) study with fathead minnows and cadmium produced flow-through LC_{50}s that were lower than comparable static LC_{50}s (~ 4,500 to 11,000 µg/L for flow-through tests versus ~30,000 µg/L for static tests). The fish used in the static tests were described as "immature" weighing about 2g (2000 mg). The size of the fish used in the Pickering and Gast (1972) their flow-through acute tests were not given, but is assumed to have been similar. In contrast, 8- to 9-day old fathead minnow fry usually weigh about 1 mg or less (EPA 2002c). Using newly hatched fry weighing about 1/1000th of the fish used by Pickering and Gast (1972) in the 1960s, cadmium LC_{50}s for fathead minnows at similar hardnesses tend to be around 50 µg/L with no obvious bias for test exposure. Similar results have been reported with brook trout. One each flow-through and static acute tests with brook trout were located, both conducted in waters of similar hardness (41 to 47 mg/L). The LC_{50} of the static test which used fry was < 1.5 µg/L whereas the LC_{50} of the flow-through test using yearlings was > 5,000 µg/L (Carroll *et al.* 1979; Holcombe *et al.* 1983).

Summary: When all other factors are equal, it appears that renewal tests may indicate chemicals are somewhat less toxic (e.g., higher LC_{50}s), but there is no clear consensus whether this indicates that renewal tests are biased toward lower toxicity than is "accurate" or whether conventional flow-through tests are biased toward higher toxicity. Comparisons with data across studies suggest that factors such as the life stage of exposures, can dwarf the influence of flow-through or renewal methods for the acute toxicity of at least metals.

2.4.1.8. The "Water-Effect Ratio" Provision

The water-quality criteria for metals proposed in this action include a Water Effects Ratio (WER) in their equations. The purpose of WERs is to empirically account for characteristics other than hardness that might affect the bioavailability and thus toxicity of metals on a site-specific basis. Because the WERs are directly incorporated into the criteria equations, no separate action is needed to change the criteria values using a WER. Following EPA's (EPA 1992) precedent, the default WER value for the proposed criteria is 1.0 "*except where the Department assigns a different value*" (Idaho Department of Environmental Quality 2011, at 210.03.c.iii.).

The concept of adjusting metals criteria to account for differences in their bioavailability in site-waters has long been a precept of water quality criteria (Carlson *et al.* 1984; EPA 1994; Bergman and Dorward-King 1997). The WER approach uses one or more standard-test species (usually *Ceriodaphnia* and/or fathead minnows) which are tested in tandem in dilution waters collected from the site of interest and in a standard reconstituted laboratory water. The results in the laboratory water are presumed to represent the types of waters used in tests used in EPA criteria documents. The WER is the ratio of the test LC_{50} in site water divided by the LC_{50} in laboratory water; the ratio is then multiplied by the aquatic life criteria to obtain a WER-adjusted site-specific criteria. The approach has probably been most used with copper because of the profound effect of DOC to ameliorate toxicity, which is not correlated with hardness.

The main problem with the concept and approach is trying to define a single "typical" laboratory dilution water that reflects that used in criteria documents. Testing laboratories may generate valid results using all sorts of different dilution waters including dechlorinated tap water, natural groundwaters (well waters), natural surface waters such as Lake Superior or Lake Erie, and reconstituted waters made from deionized water with added salts. The widely used "Interim Guidance on Determination and Use of Water-effect Ratios for Metals" (Stephan *et al.* 1994b) specified using recipes from EPA or ASTM for making standardized water that results in a water hardness with unusually low calcium relative to magnesium concentrations compared to that of most natural waters ("hardness" is the sum of equivalent concentrations of calcium (Ca) and magnesium (Mg) and is discussed more in Section 2.4.2, "The Influence of Hardness on Metals Toxicity"). This has the effect of making metals in the reconstituted laboratory waters made by standard recipe more toxic than would be expected in waters with more natural proportions of calcium and magnesium. This is because at least for fish and some invertebrates and copper, calcium reduces toxicity somewhat but magnesium affords little or no protection (Welsh *et al.* 2000a; Naddy *et al.* 2002; Borgmann *et al.* 2005b).

The effect of this issue is that unrepresentative lab waters can generate low EC_{50} values which when used as a denominator with higher EC_{50}s from site waters can produce extremely high-biased values. For instance, in WER testing on the Boise River, Idaho, a stream receiving treated municipal wastewater effluent, testing with *Ceriodaphnia* and copper resulted in mean site:lab WER of 18.4, which when multiplied by the copper CMC at a hardness of 40 mg/L would result in a WER adjusted CMC of 132 µg/L. Yet the *Ceriodaphnia* EC_{50}s in that same site water ranged from 18.6 to 60 µg/L (CH2M Hill 2002). Thus, the published WER procedure would generate a site-specific acute copper criterion that was three to seven times higher than concentration that killed 50% of a sensitive species in that same site water. Such a grossly unprotective site-specific criteria was argued for on the grounds that it was procedurally in accordance with the Idaho metals criteria under consultation, because it follows from the WER equation and definition in the NTR and derivative Idaho criteria. Because it arguably followed EPA's 1994 Interim Guidelines for developing Water Effect Ratios (Stephan *et al.* 1994b), whatever the outcome was, was therefore procedurally acceptable.

Both EPA and IDEQ have made steps to reduce the bias that could be introduced by low EC_{50} values in laboratory waters compared with site waters. The EPA (2001a) effectively eliminated the issue by setting the WER as the lesser of the site water EC_{50}/ lab water EC_{50} ratios or the ratio of site water EC_{50} divided by the SMAV from an updated criteria dataset. When this latter calculation was applied to the Boise River dataset, it produced an average copper WER of 2.6 instead of 18.4 and produced a site-specific acute copper criterion of 18.5 µg/L for a hardness of 40 mg/L (CH2M Hill 2002). Given the *Ceriodaphnia* EC_{50}s of 18.6 to 60 µg/L in site water, this approach may not fully protect species as sensitive as *Ceriodaphnia* but it's an improvement. The IDEQ (2007a) regulations at subsection 210.03.c.iii specify that calcium and magnesium ratios should be similar to those in EPA's criteria laboratory waters or the water body for which WERs are to be applied. However, such an approach was used in the Boise River project and exorbitantly high WERs still resulted so it is not clear that the WER approach can be corrected in this way. Further, IDEQ's implementation procedures for NPDES permits call specifically for the use of EPA's 1994 interim procedures (IDEQ 2007a, at subsection 210.04) although IDEQ has the discretion to use "other scientifically defensible methods" as they see fit.

Other approaches by EPA that might be used as an interim, operational substitute include establishing criteria on a more mechanistic basis that can directly account for the factors that affect toxicity. One example is the biotic ligand model (BLM) which is supposed to capture the major interactions between metals concentrations, competition, and complexation that control bioavailability and thus toxicity (Di Toro *et al*. 2001; Niyogi and Wood 2004). For copper, BLM was used as the basis of EPA's (2007a) updated aquatic life criterion, which for copper at least, should negate much of the need for empirical WER testing. The predictiveness of the copper BLM over a wide range of environmental conditions makes the BLM a more versatile and effective tool for deriving site-specific water quality criteria compared to the WER method (EPA 2000c; Di Toro *et al*. 2001).

This provision has rarely been used in Idaho, but NMFS is recommending a term and condition to help reduce future risk if WERs are developed in critical habitat for listed salmon and steelhead.

Summary: While seldom used to date, the WER is a fundamental part of the formula-based water quality criteria for metals. In guidance and practice, the manner in which WERs are developed has a substantial risk of undermining the protectiveness of criteria. Procedures that are consistent with the action evaluated in this opinion could result in criteria concentrations that were higher than concentrations that were acutely toxic to sensitive organisms when tested in the same site water. Two alternate procedures could achieve the intent of the WER provision (to adjust criteria based on site-specific conditions). First, the WER could be calculated by using the lower ratio from either (a) the site water EC_{50}/ lab water EC_{50} ratios or (b) the ratio of site water EC_{50} divided by the species mean acute value (SMAV) for that test organism (e.g., Ceriodaphnia dubia, fathead minnow, or rainbow trout) from a criterion dataset as described by EPA (2001a). Second, with copper the EPA (2007) BLM-based criteria is intended to adjust for site-specific water quality differences (EPA 2007a; DiToro *et al.* 2001).

2.4.1.9. Issue of Basing Criteria on Dissolved or Total-Recoverable Metals

One difference between the proposed action and the NTR as first published by EPA (1992) is that the proposed metals criteria are defined on the basis of "dissolved" metals rather than for "total recoverable" metals. "Dissolved" metals are those that pass through a 0.45 µm filter, and "total recoverable" metals are determined from unfiltered samples, and thus consist of both dissolved and particulate or colloidal phases. Metals sorbed to particulates are subject to gravity and will eventually settle from undisturbed water whereas dissolved metals are truly in solution and will not settle from gravity.

This criteria change was based on a 1993 EPA policy statement that "*it is now the policy of the Office of Water that the use of dissolved metal to set and measure compliance with WQSis the recommended approach, because dissolved metal more closely approximates the bioavailable fraction of metal in the water column than does total recoverable metal. This conclusion regarding metals bioavailability is supported by a majority of the scientific community within and outside the Agency. One reason is that a primary mechanism for water column toxicity is adsorption at the gill surface which requires metals to be in the dissolved form.*" (Prothro 1993).

To implement Prothro's (1993) policy change, metals criteria had to be recalculated on a dissolved basis. Because the tests in the acute and chronic datasets used to derive metals criteria were mostly reported total recoverable rather than dissolved metals, in order express metals criteria on a dissolved metals basis, a conversion was needed. To do so, Stephan (1995) evaluated what data were available on the proportions of dissolved versus total recoverable metals in different laboratories that contributed data used in the EPA metals criteria. The resulting conversion factors ranged from 0.32 with chromium (III) to 0.99 with chronic zinc. With lead, because its solubility usually decreases as hardness increases, the conversion factor for lead varies with hardness, ranging from 1 at hardness 25 mg/L to 0.69 at hardness 200 mg/L. For most metals, the conversion factors were close to 1 indicating that for the laboratory conditions under which the toxicity tests in the datasets were conducted, almost all metals were present in dissolved form (Stephan 1995)

Because no supporting documentation was given by Prothro (1993) in support of their conclusions, they are hard to evaluate. There is theoretical support for the assumption that metals need to be in dissolved form to adsorb to the gill surface (Wood *et al*. 1997), and it does seem logical to assume that metals bound to particulates would be less toxic. However, no compelling evidence was found that particulate bound metals can be assumed to be non-toxic. Only two studies were located that examined the toxicity of particulate metals in controlled experimental studies. Both found toxicity associated with particulate bound copper (Brown *et al*. 1974; Erickson *et al*. 1996).

Erickson *et al*. (1996) estimated that the adsorbed copper has a relative toxicity of almost half that of dissolved copper, and noted that the assumption that toxicity can be simply related to dissolved copper was questionable, and a contribution of adsorbed copper to toxicity cannot be generally dismissed (Erickson *et al*. 1996). One possible reason for the observed toxicity from particulate-bound copper is that adsorbed metals could become desorbed, becoming more bioavailable, as the pH of water moving across fish gills decreases. If the pH of water where a fish is living is 6 or greater, then the pH will be lowered as water crosses the gill (Playle and Wood 1989). Most ambient waters in the Snake River basin action area have pH greater than 6.

A further manner in which particulate bound metals could become biologically active is through sediment or food exposure. For instance, in Panther Creek, a tributary to the Salmon River, Idaho, total copper concentrations were measured at greater than twice that of dissolved concentrations (Maest *et al*. 1995). Copper was also greatly elevated in biofilms (algae and detritus) and sediment, and correlations between copper concentrations in benthic invertebrates and biofilms were stronger than were correlations between invertebrates and water or sediment (Beltman *et al*. 1999). Copper sorbed to sediments was also bioavailable and toxic to benthic invertebrates when exposed to Panther Creek sediments after the sediments were transferred to clean overlying water (Mebane 2002a). In this stream at the time of those studies, dissolved copper consistently exceeded dissolved criteria values, so these studies do not directly help with the question of whether streams with low contamination that largely comply with dissolved criteria could result in sediment contamination at hazardous concentrations. Others have reported toxicity from metals contaminated freshwater sediments even when overlying waters mostly are at dissolved criteria (Canfield *et al*. 1994; Besser *et al*. 2008).

Attempting to define, evaluate, and manage risks associated with contaminated sediments by basing criteria on total recoverable metals would likely be so indirect as to be ineffective. However, in the absence of such efforts the assumption that metals sorbed to particles are in effect biologically inert and can safely be ignored is questionable. The effect of this stance is to give up some conservatism in aquatic life criteria for metals.

<u>Summary:</u> The component of the action to define metals criteria as applying only to the dissolved fraction of metals rests on the rationale that metal particulates are less toxic than dissolved metals. Criteria are adjusted from total to dissolved metals fraction through conversion factors. The total to dissolved conversion factors for metals criteria were set in a generally conservative manner and are close to 1 for most metals. While the conversion factors per se are not a conservation problem, the concept of basing criteria solely on the dissolved fraction may not always be protective. While we concur that for divalent metals (e.g., cadmium, copper, lead, nickel, zinc), the particulate fraction is less toxic, the particulate fraction is not necessarily non-toxic. Conceptually, the particulate fractions of metals and inorganics could contribute to foodweb exposure pathways from sediments or biofilms to macroinvertebrates to fish. This is of particular concern for substances with primarily dietary routes of exposure (e.g., arsenic, mercury, and selenium).

2.4.1.10. Mixture Toxicity: criteria were developed as if exposures to chemicals occur one at a time, but chemicals always occur as mixtures in effluents and ambient waters

In point or nonpoint pollution, chemicals occur together in mixtures, but criteria for those chemicals are developed in isolation, without regard to additive toxicity or other chemical or biological interactions (Table 2.4.1.1). Whether the toxicity of chemicals in mixtures is likely greater or less than that expected of the same concentrations of the same chemicals singly is a complex and difficult problem. While long recognized, the "mixture toxicity" problem is far from being resolved. Even the terminology for describing mixture toxicity is dense and has been inconsistently used (e.g., Sprague 1970; Marking 1985; Borgert 2004; Vijver *et al*. 2010). One scheme for describing the toxicity of chemicals in mixtures is whether the substances show additive, less than additive, or more than additive toxicity. The latter terms are roughly similar to the terms "antagonism" and "synergism" that are commonly, but inconsistently used in the technical literature.

For both metals and organic contaminants that have similar mechanisms of toxicity (e.g., different metals, different chlorinated phenols), assuming chemical mixtures to have additive toxicity has been considered a reasonable and usually protective (Norwood *et al*. 2003; Meador 2006). This conclusion is in conflict with the way effluent limits are calculated for discharge of toxic chemicals into receiving water. Each projected effluent chemical concentration occurring during design flow is divided by its respective criterion, along with adjustments for variability and mixing zone allowances (EPA 1991). Thus, each substance would be allowed to reach one "concentration unit" and any given discharge or cleanup scenario would likely have several concentration units allowed, which is sometime referred to as cumulative criterion units.

Experimental approaches in the literature usually report "toxic units" (TUs) based on observed toxicity in single substance tests, rather than criterion units. In this "concentration addition" scheme, toxicity of different chemicals is additive if the concentrations and responses can be summed on the basis of "TUs." For instance, assume for simplicity that cadmium is more toxic than copper to a species, with the an EC_{50} of 4 µg/L for cadmium, and an EC_{50} of 8 µg/L for copper. We will also call each single metal EC_{50} a TU. The toxicity of mixtures could be estimated as follows:

$$4\ \mu g/L\ Cd + 0\ \mu g/L\ Cu = \frac{4\ \mu g/L}{4\ \mu g/L/TU} + \frac{0\ \mu g/L}{8 \mu g/L/TU} = 1\ TU$$, (obviously, for a single substance), *or*

$$2\ \mu g/L\ Cd + 4\ \mu g/L\ Cu = \frac{2\ \mu g/L}{4\ \mu g/L/TU} + \frac{4 \mu g/L}{8 \mu g/L/TU} = 0.5 + 0.5 = 1\ TU.$$

Using this approach, some studies have shown significant additive toxicity. For instance, Spehar and Fiandt (1986) exposed rainbow trout and *Ceriodaphnia dubia* simultaneously to a mixture of five metals and arsenic, each at their acute CMC, which by definition were intended to be protective. There were no survivors. In chronic tests, adverse effects were observed at mixture concentrations of one-half to one-third the approximate chronic toxicity threshold of fathead minnows and daphnids, respectively, suggesting that components of mixtures at or below no effect concentrations may contribute significantly to the toxicity of a mixture on a chronic basis (Spehar and Fiandt 1986).

A common outcome in metals mixture testing has been that metals combinations have been less toxic than the sum of their single-metal toxicities, i.e., show less than additive toxicity or are antagonistic (Finlayson and Verrue 1982; Hansen *et al*. 2002c; Norwood *et al*. 2003; Vijver *et al*. 2011; Mebane *et al*. 2012). The other possibility, more than additive toxicity (also called synergistic effects) are rare with metals although it has been shown with pesticides (Norwood *et al*. 2003; Laetz *et al*. 2009).

<u>Summary:</u> The water criteria evaluated in this opinion were all developed as if only one chemical was present at a time. However, in the real world chemicals always occur in mixtures. As result, criteria and discharge permits based upon them may afford less protection than intended. Measures to address this potential underprotection need to be included in discharge permits.

The efficacy of whole-effluent toxicity tests to evaluate mixture toxicity. The EPA's approach to the mixture toxicity problem in effluents, including effects of substances without numeric criteria or unmeasured substances, has been to recommend an integrated approach to toxics control (EPA 1991, 1994). The EPA has long recognized that numerical water quality criteria are an incomplete approach to protecting or restoring the integrity of water. A major part of EPA's strategy for measuring and controlling such potential issues has been through the concept of an integrated approach to toxics control, where meeting numerical criteria is but one of three elements. The other two elements are: (1) The concept of regulating whole effluents through whole- effluent toxicity (WET) testing; and (2) through biological monitoring of ambient waters that receive point or nonpoint discharges (EPA 1991, 1994). Because of assumptions that: (1) Chemicals will inevitably occur in ambient waters in mixtures rather than occurring chemical by chemical in the fashion that criteria are developed; and (2) it's not possible to know all the potential contaminants of concern in effluents and receiving waters, let alone measure them, it is not feasible to predict effects by chemical concentrations alone. Thus, the EPA developed procedures for testing the whole-toxicity of effluents and receiving waters, including procedures for identifying and reducing toxicity (e.g., Mount and Norberg-King 1983; Norberg-King 1989; Mount and Hockett 2000). In practice, some consideration of the potential for aggregate toxicity through WET testing is made by EPA for major permits that they administer in Idaho.

Test procedures for WET testing are intended to be practical for permitted dischargers or test laboratories to carry out as a routine monitoring tool. Thus, to simplify testing, improve test repeatability, and to facilitate interpretation of test results by dischargers and permit compliance staff, the EPA has limited WET testing requirements to select standard test species and test conditions (EPA 2002a, 2002c). Most commonly, EPA has required monitoring for chronic WET through testing of two species, fathead minnows and the cladoceran ("water flea") *Ceriodaphnia dubia*. Both tests are administered as 7-day tests. *Ceriodaphnia* have a short life-cycle, so even though the test is only 7 days, it spans three broods, and so can be considered a "true" chronic test that includes all or most of an organism's life cycle. In contrast, the 7-day fathead minnow "chronic" test only spans about 1% of the 2-year or so life span of a fathead minnow and is more properly called a short-term method for predicting chronic toxicity.

The rationale and performance of WET testing for predicting or protecting against impairment have been complicated and controversial and have been debated in conferences and articles, among them a special issue of the journal *Environmental Toxicology and Chemistry* (v19, 1, January 2000) and an entire book (Grothe *et al*. 1996). Issues with WET testing include whether the tests are sensitive, and whether any single species toxicity test can meaningfully predict in stream effects or lack thereof. For instance, Clements and Kiffney (1996) noted that *Ceriodaphnia* effluent tests were correlated with effects detected from stream microcosms or field surveys, but the latter two tended to be more sensitive than the Ceriodaphnia effluent tests. Conversely, Diamond and Daley (2000) and de Vlaming *et al*. (2000) found that the chronic WET methods were useful for predicting ambient impairment.

The best comparison of the sensitivity of WET tests in relation to listed salmon, steelhead and their prey is probably a series of tests conducted at the same laboratory with the same dilution water with copper and different species (Table 2.4.1.2). Neither the *Ceriodaphnia* or 7-day fathead minnow test were as sensitive as 30- or 60-day chronic tests with rainbow trout; the

Ceriodaphnia were about twice as resistant as the rainbow trout, and the 7-day fathead minnow test was almost five times as resistant as the longer rainbow trout test. Dwyer *et al.* (2005a) also found that the *Ceriodaphnia* test was considerably more sensitive than the 7-day fathead test to a complex "effluent" comprised of a mixture of pesticides, chlorinated organic compounds, ammonia, and metals. The low sensitivity of the 7-day fathead minnow test might be because the species is inherently less sensitive to some substances than salmonids or because a 7-day exposure is too short to be an accurate "short-term" chronic measurement (Suter 1990; Lazorchak and Smith 2007).

Comparisons with other metals were less reliable because they required comparing tests across studies and regression-based hardness normalizations (Table 2.4.1.3). Focusing on the more sensitive *Ceriodaphnia* test, sensitivity comparisons were made for four metals with rainbow trout (treating rainbow trout as a surrogate for listed salmon and steelhead). The comparisons used the most convenient, readily available statistics that were comparable across tests, even though those statistics do not reflect protective concentrations in of themselves (e.g. EC20, MATC, see "*Implications of the use of the "chronic value" statistic*"). A sensitivity ratio of 1.0 or less suggests that *Ceriodaphnia* are at least as sensitive as the salmonid surrogate and that the WET testing should be protective for aggregate, direct toxicity of waste mixtures in effluents (Table 2.4.1.2). The comparisons suggest that for cadmium and zinc the *Ceriodaphnia* test would be almost as sensitive or more sensitive as the average rainbow trout test; however, for copper and lead. Chinook salmon or rainbow trout could be much more sensitive than the *Ceriodaphnia*.

A further consideration beyond these simple comparisons of whether reduced survival or reproduction in Ceriodaphnia test results occurred at higher or lower concentrations than mortality to listed salmonids, is whether WET tests such as Ceriodaphnia can be used as a proxy indicator of sublethal effects of chemicals to salmonids, such as olfactory impairment. The limited information available suggests that they can be used in this way, at least for copper. Toxicity of copper to aquatic organisms can often be predicted using a "biotic ligand model" or BLM. The BLM uses geochemical speciation modeling to model bioaccumulation of copper on the organisms' gills or their other biological tissues in contact with water (i.e., their "biotic ligands"), and then uses an empirical species-specific toxicity adjustment to predict effects (Appendix C). This empirical species-specific toxicity adjustment was initially done to predict killing organisms with different sensitivities following short-term exposures (EPA 2007a). However, it has been successfully expanded to predict olfactory impairment (or lack thereof) in coho salmon or behavioral avoidance in rainbow trout or Chinook salmon (Appendix C; Meyer and Adams 2010). These analyses suggest that on the average, adverse effects predicted for Ceriodaphnia dubia would occur at lower copper concentrations than would olfactory impairment or avoidance behavior in rainbow trout, based upon lower modeled critical accumulation values for Ceriodaphnia dubia (0.06 vs. 0.19 nmol/g wet weight (Appendix C; Meyer and Adams 2010).

In contrast, the Ceriodaphnia WET test has been shown to be able to predict adverse effects in benthic macroinvertebrate communities in streams, but that the Ceriodaphnia WET test appeared less sensitive than the more complex stream communities (Clements and Kiffney 1996). This suggests that with a sensitivity adjustment, the Ceriodaphnia WET test could be used to predict whether effluents were likely to adversely modify critical habitats by reducing the benthic macroinvertebrate forage base for rearing salmonids.

Table 2.4.1.2. Relative sensitivity of standard 7-day WET tests with *Ceriodaphnia* and fathead minnows to rainbow trout with copper under directly comparable test conditions (ASTM moderately-hard water, hardness 170 mg/L).

Organism	Test duration	EC25 for the most sensitive endpoints (µg/L)	Source
Rainbow trout	30-days (starting with fry)	21	(Besser *et al.* 2005b)
Rainbow trout	60-days (starting with eggs)	25	(Besser *et al.* 2005b)
Fathead minnow	30-days	12-24 (range of 3 replicate tests)	(Besser *et al.* 2005b)
Fathead minnow	7-days	103	(Dwyer *et al.* 2005a)
Ceriodaphnia dubia	7-days	51	(Dwyer *et al.* 2005a)

Table 2.4.1.3. Relative sensitivity of the standard WET *Ceriodaphnia dubia* 7-day test in relation to a surrogate salmonid for listed salmon and steelhead (rainbow trout except where noted), pooled from data compilations

Metal	*Ceriodaphnia dubia* SMCV (µg/L)	Surrogate salmonid SMCV (µg/L)	Sensitivity Ratio (*C.dubia* ÷ Salmonid)	Notes (source)
Cd	2.04	1.7	1.2	MATC, (Mebane 2006)
Cu	19	23.8	0.8	EC20s, (EPA 2007a);
Cu	19	5.9	3.2	Chinook salmon biomass EC20 (EPA 2007a); Rainbow trout, geometric mean of 5 tests, normalized to hardness 50; (Mebane *et al.* 2008); C. dubia is from a single test at hardness
Pb	46	28	1.6	52 mg/L, pH 7.56 (Mager *et al.* 2011a) (note)
Zn	33	113	0.3	NOECs; (Van Sprang *et al.* 2004)

Note: Much new data with C. dubia and chronic toxicity of Pb has been recently generated (Parametrix 2010; Mager *et al.* 2011a). While this was too much to synthesize and estimate whether C. dubia are usually more or less than salmonids, recent toxicity values with C. dubia indicate the sensitivities overlap those of rainbow trout and the species may be much more sensitive than previously indicated (Jop *et al.* 1995; Mebane *et al.* 2008)

Summary: Our review generally supports EPA's concept of assessing mixture toxicity of criteria substances under consultation through WET testing and instream bioassessment. However, the more sensitive of the two commonly used chronic WET tests, the three-brood *Ceriodaphnia dubia* test was sometimes less sensitive than chronic tests with salmonids. The 7-day fathead minnow test was consistently less sensitive than chronic salmonid tests in the data reviewed. This suggests that to be protective of listed salmonids, the assessment triggers for the

Ceriodaphnia test might have to be scaled to account for sensitivity and or differences in tolerable risk for a threatened species versus a zooplankton.

In much of EPA's (2000a) biological evaluation of the action, and elsewhere in the present opinion, the effects of criteria provisions or substances are evaluated linearly, one-by-one. Despite this simplification, in the environment chemicals in water never occur in isolation, but rather always occur as mixtures. The toxicity of mixtures is probably dependent upon many factors, such as which chemicals are most abundant, their concentration ratios, differing factors affecting bioavailability, and organism differences. Because of this complexity, accurate predictions of the combined effects of chemicals in mixtures appear to be beyond the present state of the ecotoxicology practice.

Here, despite the complexities and many exceptions, we make a general assumption that, at their criteria concentrations, the effects of chemicals in mixtures would likely be more severe than would be the same concentration of the mixture components singly.

Addressing mixture toxicity through the use of WET testing and instream bioassessment are practical and reasonable approaches for addressing the expected increased toxicity of a given concentration of a chemical in the presence of other chemicals. However, the assessment triggers on WET tests may not be sensitive enough to protect listed salmonids with reasonable certainty, and biomonitoring has not always been well defined. Measures for implementing biomonitoring are provided in Section 2.9 and Appendix E

2.4.1.11. Frequency, Duration and Magnitude of Allowable Criteria Concentration Exposure Exceedences.

For simplicity, much of the discussion of the water quality criteria that are the subject of this consultation treats the criteria as though they were defined solely as a concentration in water. However, the action actually defines aquatic life criteria in three parts: a concentration(s), a duration of exposure, and an allowable exceedence frequency. All of EPA's criteria recommendations define criteria using a statement similar to the following:

> "The procedures described in the 'Guidelines for Deriving Numerical National Water Quality Criteria for the Protection of Aquatic Organisms and their uses' indicate that, except possibly where a locally important species is very sensitive, freshwater aquatic organisms and their uses should not be affected unacceptably if the 4-day average concentrations of [the chemical] do not exceed [the 'chronic' criterion continuous concentration] more than once every 3 years on the average and if the 1-hour average concentration does not exceed [the 'acute' criterion maximum concentration] more than once every 3 years on the average."

The 4-day and 1-hour duration and averaging periods for criteria were based upon judgments by EPA authors that included considerations of the relative toxicity of chemicals in fluctuating or constant exposures. The EPA's (1985) Guidelines considered an averaging period of 1 hour most appropriate to use with the criterion maximum concentration or (CMC or "acute" criterion) because high concentrations of some materials could cause death in 1 to 3 hours. Also, even when organisms do not die within the first few hours, few toxicity tests attempt to monitor for latent mortality by transferring the test organism into clean water for observation after the chemical exposure period is over. Thus, it was not considered appropriate to allow concentrations above the CMC for more than 1 hour (Stephan *et al.* 1985). Recent criteria documents (e.g., EPA 2007a) have used an averaging period of 24 hours for their CMC, although no explanation could be found for the deviation from the 1985 Guidelines and thus, the issue of latent toxicity might not have been considered.

A review of more recent information supported EPA's judgments from the 1980s that if an averaging period is used with acute criteria for metals, it should be short. Some of the more relevant research relates the rapid accumulation of metals on the gill surfaces of fish to their later dying. When fish are exposed to metals such as cadmium, copper, or zinc, a relatively rapid increase in the amount of metal bound to the gill occurs above background levels. This rapid increase occurs during exposures on the order of minutes to hours, and these brief exposures have been sufficient to predict toxicity at 96 to 120 hours. The half saturation times for cadmium and copper to bind to the gills of rainbow trout may be on the order of 150 to 200 seconds (Reid and McDonald 1991). Several other studies have shown that exposures well under 24 hours are sufficient for accumulation to develop that is sufficient to cause later toxicity (Playle *et al.* 1992; Playle *et al.* 1993; Zia and McDonald 1994; Playle 1998; MacRae *et al.* 1999; Di Toro *et al.* 2001). Acute exposures of 24 hours might not result in immediate toxicity, but deaths could result over the next few days. Simple examination of the time-to-death in 48- or 96-hour exposures would not detect latent toxicity from early in the exposures. The few known studies that tested for latent toxicity following short-term exposures have demonstrated delayed mortality following exposures on the order of 3 to 6 hours (Marr *et al.* 1995a; Zhao and Newman 2004, 2005; Diamond *et al.* 2006; Meyer *et al.* 2007a). Observations or predictions of appreciable mortality resulting from metals exposures on the order of only 3 to 6 hours supports the earlier recommendations by Stephan and others (1985) that the appropriate averaging periods for the CMC is on the order of 1 hour.

The 4-day averaging period for chronic criteria was selected for use by EPA with the CCC for two reasons (Stephan *et al.* 1985). First, "chronic" responses with some substances and species may not really be due to long-term stress or accumulation, but rather the test was simply long enough that a briefly occurring sensitive stage of development was included in the exposure (e.g., Chapman 1978a; Barata and Baird 2000; De Schamphelaere and Janssen 2004; Grosell *et al.* 2006b; Mebane *et al.* 2008). Second, a much longer averaging period, such as 1 month would allow for substantial fluctuations above the CCC. Whether fluctuating concentrations would result in increased or decreased adverse effects from those expected in constant exposures seems to defy generalization. A comparison of the effects of the same average concentrations of copper on developing steelhead, *Oncorhynchus mykiss*, that were exposed either through constant or fluctuating concentrations found that steelhead were about twice as resistant to the constant exposures as they were to the fluctuating exposures (Seim *et al.* 1984). Similarly, *Daphnia magna* exposed to daily pulses of copper for 6 hours at close to their 48-hour LC_{50} concentrations had more severe effects after 70 days than did comparisons that were exposed to constant copper concentrations that were similar to the average of the daily fluctuations (Ingersoll and Winner 1982). In contrast, cutthroat trout exposed instream to naturally fluctuating zinc concentrations survived better than fish tested under the same average, but constant zinc concentrations (Nimick *et al.* 2007; Balistrieri *et al.* 2012). Thus, literature reviewed either supports or at least do not contradict EPA's position on averaging periods.

The third component of criteria, EPA's once-per-3-years allowable exceedence policy was based on a review of case studies of recovery times of aquatic populations and communities from locally severe disturbances such as spills, fish eradication attempts, or habitat disturbances (Yount and Niemi 1990; Detenbeck *et al*. 1992). In most cases, once the cause of the disturbance was lifted, recovery of populations and communities occurred on a time frame of less than 3 years. The EPA has subsequently further evaluated the issue of allowable frequency of exceedences through extensive mathematical simulations of chemical exposures and population recovery. Unlike the case studies, these simulations addressed mostly less severe disturbances that were considered more likely to occur without violating criteria (Delos 2008). Unless the magnitude of disturbance was extreme or persistent, this 3-year period seemed reasonably supported or at least was not contradicted by the information we reviewed.

A more difficult evaluation is the exceedence magnitude, which is undefined and thus not limited by the letter of the criteria. Thus, by the definition, a once-per-3-year exceedence that has no defined limits to its magnitude, could be very large, and have large adverse effects on listed species. However, within the 4-day and 1-hour duration constraints of the criteria definitions, some estimates of the potential magnitude of exceedences that could occur without "tripping" the duration constraints can be calculated. This is because environmental data such as chemical concentrations in water are not unpredictable but can be described with statistical distributions, and statements of exceedence probabilities can be made. Commonly with water chemical data and other environmental data, the statistical distributions do not follow the common bell-curve or normal distribution, but have a skewed distribution with more low than high values. This pattern may be approximated with a log-normal statistical distribution (Blackwood 1992; Limpert *et al.* 2001; Helsel and Hirsch 2002; Delos 2008).

The following three hypothetical scenarios are intended to illustrate contaminant concentrations that could occur without violating the exceedence frequency and duration limitations of the proposed criteria (Figure 2.4.1.7). The scenarios use randomly generated values from a log-normal distribution with different variabilities and serial correlations. Serial correlation refers to the pattern in environmental data where values at time one are often highly correlated with values at time two and so on. For example, a hot day in summer is much more likely to be followed by another hot day than a bitterly cold day, a low chemical concentration during stable low flows on a day in September will most likely be followed by low chemical concentration the next day, a high chemical concentration in a stream during runoff on a day in April will more likely to be repeated by another high concentration, and so on (Helsel and Hirsch 2002; Delos 2008). Under Scenario 1, effects could be appreciable since the mean concentrations are close to the criteria, and organisms would have little relaxation of exposure for recovery. Under Scenario 2, effects to a population of sensitive organisms would presumably be slight, since the mean concentrations were well below the criterion, and the exceedence magnitude was slight followed by a recovery opportunity. Scenario 3 might be more likely in runoff of nonpoint pollutants from snowmelt or stormwater. In these scenarios, sensitive populations could experience effects ranging from appreciable reductions if the contaminant pulse hit during a sensitive part of their life history, to no effect if it hit during a resistant phase or if the listed species was less sensitive than the species that drove the criteria calculations.

An actual event that was very similar to Scenario 3 occurred when an upset at a large, industrial mining operation caused elevated cadmium concentrations in Thompson Creek, a tributary to the upper Salmon River in Idaho. In April 1999, a pulse of cadmium about 30X higher than background, 2.6 times higher the chronic criterion, and equal to the acute criterion was detected. The duration of exceedence was probably greater than a day and less than a week. By August 1999, when a biological survey was conducted, few if any adverse effects could be detected in the benthic community structure. Whether subtle differences between unaffected upstream survey sites were lingering effects of the disturbance or just differences in naturally patchy stream invertebrate communities was unclear. However, it does suggest that benthic communities in similar mountain streams would be either resilient to, or recover quickly from criteria exceedences of this magnitude (Mebane 2006, pp. 47,62).

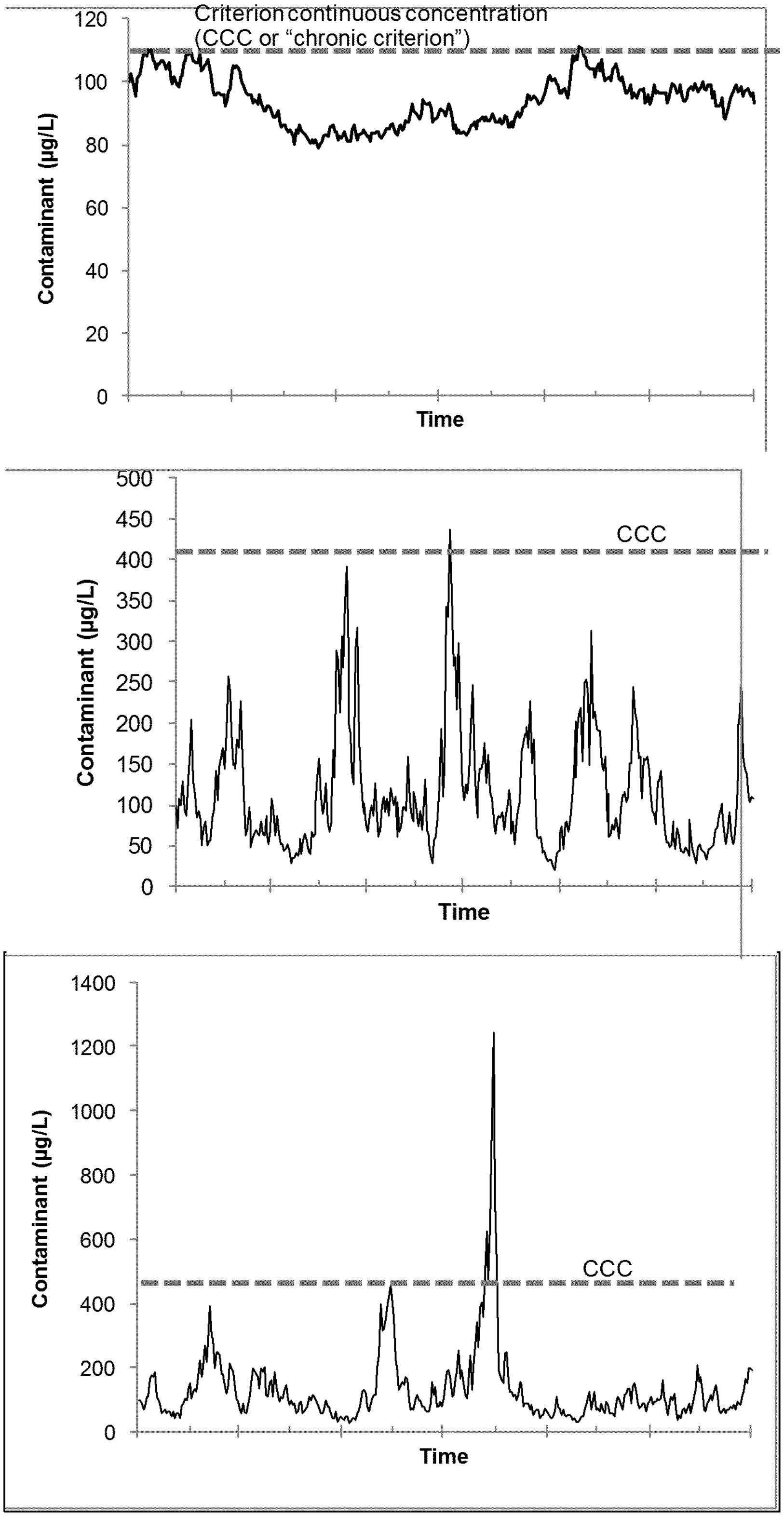
These hypothetical scenarios used a simplified, fixed criterion, whereas in actuality, some of EPA's criteria vary and may be positively correlated with the concentrations of metals in water. If the criteria accurately reflect risks from varying environmental conditions, and if ambient conditions co-vary with and are positively correlated with criteria, this will tend to lessen risks resulting from ambient increases in concentration. In cases where the criteria were positively correlated with the contaminants, such as in the following Section 2.4.4 example for Pine Creek with cadmium or the BLM-copper example for Panther Creek, the frequency and magnitude of exceedences is expected to be less than if the criteria and contaminant concentrations did not rise and fall together. This is because the contaminant and another water quality parameter that mitigates toxicity have common sources and rise and fall together, such as cadmium and calcium in Pine Creek where the source for both is probably weathering of gangue rock and spring snowmelt and runoff appears to dilute both.

In the Panther Creek example, copper and DOC tended to rise and fall together with snowmelt and runoff, similarly mitigating exceedence frequency and magnitude. This was the case in all examples examined. In the Panther Creek example, the hardness-based criterion is negatively correlated with copper concentrations, which gives the impression of risks of copper being exacerbated due to lower hardness corresponding with higher copper. However, this impression is probably misleading because copper risks indicated from the hardness-based criteria are often the opposite from risks indicated by BLM-based criteria, which is considered to more accurately represent the copper risks (Section 2.4.4; Appendix C).

While NMFS did not locate any plausible examples of negative correlations between contaminants and important factors modifying toxicity, it is likely that such scenarios do occur somewhere because if the event that releases the contaminant, such as a runoff pulse from a storm or snowmelt, caused a contaminant spike from washing accumulations into a stream and at the same time lowered the pH and hardness, then the magnitude of exceedences could be more severe. Such a circumstance could be plausible for metals such as cadmium, lead, or zinc in which hardness is a major modifier of toxicity.

Further, the actual possibility that an extreme exceedence would occur and be "allowed" under the exceedence policy seems unlikely. This is because in natural waters seasonal and hydrologic factors tend to cause concentrations to be serially correlated, that is low concentrations follow low concentrations and high concentrations follow high concentrations (Helsel and Hirsch 2002;

Delos 2008). Thus for an extreme exceedence to be allowable under the chronic criteria 4-day average concentration definition, it would also have to not exceed the 1-hour acute criteria definition. A very large exceedence of the sort illustrated in Figure 2.4.1.7, Scenario 3, would likely span across more than one, 1-hour averaging period for acute criteria and "violate" the one exceedence per 3-year recurrence interval term. While there are no regulatory limits on the upper concentration of an exceedence of the 1-hour acute criteria, the idea that a chemical concentration in a natural water could rapidly rise to acutely toxic concentrations and then drop back down to below criteria seems like a remote possibility. In urban watersheds with high proportions of impervious surface, runoff is flashier than in forested watersheds, and short-term pulse exposures could occur in those settings Booth *et al.* (2002). In the predominately forested areas of the action areas, such scenarios seem less likely.



Scenario 1: Contaminant concentrations have low variability, and while the CCC is only briefly exceeded, the average exposure concentration is only slightly lower than the criterion. Such a scenario might result from a stable effluent discharged into a flow regulated receiving water .

Scenario 2: Contaminant concentrations are more variable, and while the frequency and magnitude of criterion exceedences are similar to scenario 1, average concentrations are well below the CCC in this scenario. Such a scenario might result from nonpoint pollutants resulting from snowmelt or precipitation into an unregulated stream, such as stormwater from a mining operation.

Scenario 3: Contaminant concentrations have the same variability as scenario 2, but by chance a high magnitude criterion exceedence of 12X above the average concentrations occurred. Unless the acute criterion for this substance was at least 12X higher than the CCC, such an exceedence would not be allowable because the 1-hour acute criterion averaging period would also be exceeded.

Figure 2.4.1.7. Three example allowable scenarios for criteria exceedence magnitudes

Summary: The 1-hour and 4-day exceedence durations for acute and chronic criteria respectively are supported by the science as reasonable and adequately protective. Whether the allowable 1 in 3 years exceedence frequency is sufficiently protective was difficult to evaluate, in part because the magnitude of allowable exceedences is undefined. However, the likelihood that a runoff pulse could both rise and fall so high within an hour that it could cause acute effects without exceeding the acute criteria seems unlikely. This does remain an aspect of uncertainty regarding the protectiveness of criteria.

2.4.1.12. Special Consideration for Evaluating the Effects of the Action on Critical Habitat

Fundamentally, the analyses of water quality criteria for toxic substances included in this Opinion are most directly analyses of the "water quality" features of the PCE's of critical habitat. The WQS directly characterize and define the conditions and quality of surface waters that listed salmon and steelhead experience, either as incubating embryos in the interstices of spawning gravels, or as juveniles and adults in the water column. Analyzing whether the action would represent an "adverse modification" of water quality is at least conceptually more straightforward than whether these modifications would jeopardize the continued existence of listed species. This is because quantitative causal predictions relating habitat change to species population changes and long-term viability are uncertain. Many simplifying assumptions are required, including things like specifics of species life histories, other interacting physical and biological factors, the nature and magnitude of assumed exposures such as whether the exposures are joint or separate, continuous or intermittent, magnitude of exceedences, and so on. Quantitative models relating water quality changes to extinction risks may provide value in a relative sense for evaluating relative risks of different "what if" scenarios (e.g., McCarthy *et al.* 2004; Baldwin *et al.* 2009; Mebane and Arthaud 2010). However, except for cases of extreme-risk with very high extinction probabilities (perhaps for example, Spromberg and Scholz 2011), the absolute projections from quantitative models of habitat and population changes may be thought of as mathematical speculation. Further, all mathematical population models will project some extinction risk, and policy definitions or scientific consensus are elusive on how much habitat modification or extinction risk is too much under narrative Endangered Species Act definitions (DeMaster *et al.* 2004; McGowan and Ryan 2009; McGowan and Ryan 2010; Owen 2012).

The types of adverse effects reported in the scientific literature that we consider to directly or indirectly reduce survival or reproduction included such things as reductions in survival, growth, swimming performance, ability to detect or evade predators (e.g., chemoreception), ability to detect or capture prey, ability to detect and avoid harmful concentrations of chemicals, homing ability, disease resistance, certain fish health indicators that have been related to survival or growth such as gill or liver tissue damage, spawning success, or fecundity. For evaluating what severity of effects to invertebrates would be considered an appreciable enough reduction in forage to reduce the conservation value of habitats for freshwater rearing, if a general reduction in diversity or abundance of invertebrates was expected at criteria conditions, we would consider that to be "appreciable." Because salmonids are opportunistic feeders, effects to a single invertebrate species for example, might not be important. This assumption must be tempered by the availability of data. Often data were available for very few invertebrate species, so if few data were available, but they indicated adverse effects, that could be considered a diminishment in water quality and habitat value.

Examples of types of effects that we do not consider to be sufficiently severe to represent an "appreciable diminishment" of water quality and thus the value of critical habitat include simple bioaccumulation of chemical in tissues, enzyme changes, gene expression or transcription, molecular changes, or other markers of exposure that may be considered sub-organismal, without known correlation to other changes such as reduced growth or survival. A human-health analogy of the latter types of effects would be those considered asymptomatic or sub-clinical, that is, not rising to the level that caused negative symptoms.

Because multiple criteria (acute and chronic aquatic life criteria, human health based water quality criteria) for the same substances would apply to any given area of critical habitat, we compared adverse effects indicated from short-term experiments of 4 days or less duration to the acute criteria that are intended to protect against short-term effects, and compared adverse effects shown in longer-term studies to the proposed chronic criteria. Human health-based criteria were only evaluated if they were both more stringent than chronic criteria and if the chronic criteria failed to be fully protective. In Idaho, water quality criteria for the protection of "fishable" beneficial uses based on avoiding health risks from consuming tainted fish, were clearly intended to be some sort of backstop to the aquatic life criteria because the human-health based criteria explicitly apply to waters designated for "cold water biota" and "salmonid spawning" aquatic life uses (Table 1.3.1).

For most of the substances, there were at least some conflicts in the scientific literature where for the same species and similar types of experiments, one study might find no ill effects from a given concentration and another might find severe effects. Thus, we considered the overall strength of the evidence for or against the protectiveness of criteria.

Sediments. If sufficiently elevated, toxic pollutants in ambient water may adversely modify critical habitat through contamination of stream and lake bed sediments. In general, sediment contamination by toxic pollutants adversely modifies critical habitat because the particulate forms of toxicants are either immediately bioavailable through re-suspension, or are a delayed source of toxicity through bioaccumulation or when water quality conditions favor dissolution at a later date. Specifically, contaminated sediments are expected to influence: (1) The intra-gravel life stages of listed salmon and steelhead; (2) the food source of listed salmonids; and (3) the fish through direct ingestion or deposition on the gill surfaces of particulate forms of toxicants. However, other than for mercury, it is not clear whether moderately-elevated concentrations in water (i.e, up to criteria concentrations), would be likely to result in concentrations in bed sediments that are elevated to a degree that would pose appreciable risks to listed salmonids or their prey.

The proposed criteria do not explicitly account for exposure to contaminants via sediments. NMFS recognizes that considerable technical and practical problems exist in defining water quality criteria on a sediment basis, and that this is presently the subject of considerable research and debate. Nevertheless, most organic and metal contaminants adsorb to organic particulates and settle out in sediments. Thus, at sites where there have been past discharges, or where there are continuing discharges of contaminants into the water column, sediments form a long-term repository and a continuing source of exposure that must be addressed if the water quality component of critical habitat is to be protected. Further, although these substances may not readily be transferred into the water column, they may still be available to salmonids through food chain transfer from their benthic prey, or through ingestion of sediment while feeding, as has been described in preceding sections. Not having water quality criteria that consider uptake through direct ingestion or food chain transfer leaves potential routes for harm to listed species that the proposed criteria do not directly address.

Salmonid Prey Items. An important type of indirect adverse effect of toxic substances to listed salmon and steelhead is the potential reduction of their invertebrate prey base. This is because for many substances, invertebrates tend to be among the most sensitive taxonomic groups and because juvenile salmonids depend on aquatic invertebrates during freshwater rearing. Known effects of specific substances to invertebrates are discussed specifically in those sections; however, some general considerations and assumptions applicable to all substances follow.

First, in instances of a pulse of chemical disturbance such as insecticide spraying of forests or crops, effects to aquatic invertebrate communities ranging from increased drift to catastrophic reductions can result (Ide 1957; Gibson and Chapman 1972; Wallace and Hynes 1975; Wallace *et al*. 1986). In such cases, even if the fish are not directly harmed by the chemical, the temporary reduction in food from the reduction in invertebrate prey can lead to reduced growth, and reduced growth in juvenile salmonids can in turn be extrapolated to reduced survival and increased risk of population extinction (Kingsbury and Kreutzweiser 1987; Davies and Cooke 1993; Baldwin *et al*. 2009; Mebane and Arthaud 2010). However, such severe effects would not be expected in waters with chemical concentrations similar to the maximum allowed by aquatic life criteria. The criteria are intended to only allow adverse effects to a small minority of the species in aquatic communities, and for most substances, the analyses of individual criteria that follow in Sections 2.4 are consistent with this expectation (*although copper has exceptions*).

This begs the question, whether the loss of a minority of invertebrate prey species could lead to a reduction in forage for juvenile salmonids that in turn could affect growth and survival? To address that question, NMFS reviewed a large number of studies on food habits of salmonids in streams, lakes, and reservoirs.[5] The body of evidence indicates that juvenile salmonids are opportunistic predators on invertebrates, and so long as suitable, invertebrate prey items are abundant and diverse, the loss of a few "menu items" probably would not result in obvious, adverse effects. Suitable invertebrate prey items for juvenile salmonids are those that are small enough to be readily captured and swallowed, and vulnerable to capture (i.e., not taxa that are burrowers or are armored (Keeley and Grant 2001; Suttle *et al*. 2004; Quinn 2005)). Some otherwise apparently suitable taxa such as water mites (Hydracarina) appear to taste bad to salmonids and others, like copepods, are too small to provide much energy for the effort it takes to eat them (Keeley and Grant 1997). Freshwater aquatic invertebrates have such great diversity (over 1200 species in Idaho alone, Mebane 2006), that they have some ecological overlap and redundancy, so that the loss of a few species would be unlikely to disrupt the stream or lake ecology greatly (Covich *et al*. 1999). However, this apparent ecological redundancy is compromised in streams that have already lost substantial diversity to pollution. For instance, in

[5] Over 90 were reviewed, although only a handful are listed here.

copper-polluted Panther Creek, Idaho, during springtime in the early 1990s, the total count of invertebrates was just as abundant as in reference sites, although the abundance was composed of fewer species. Yet in October, the abundance in the polluted reaches was less than 10% of reference (Mebane 1994). With reduced diversity, after a single species hatches and leaves the streams, a large drop in remaining abundance can occur. Because all species don't hatch at the same time, with greater diversity, the swings in abundance would be less severe. Further, in copper-polluted tributaries to Panther Creek, the usually abundant mayflies were scarce and had been replaced by unpalatable mites and low-calorie copepods (Todd 2008).

One consistent theme in the literature on the feeding of salmonids in streams is the persistent importance of mayflies and chironomid midges (Chapman and Quistorff 1938; Chapman and Bjornn 1969; Sagar and Glova 1987, 1988; Mullan *et al.* 1992; Clements and Rees 1997; Rader 1997; White and Harvey 2007; Iwasaki *et al.* 2009; Syrjänen *et al.* 2011). In lakes zooplankton are disproportionally important, and as stream size increases and gradients drop, amphipods become popular food items with migrating and rearing juvenile salmon and steelhead (Tippets and Moyle 1978; Rondorf *et al.* 1990; Muir and Coley 1996; Budy *et al.* 1998; Karchesky and Bennett 1999; Steinhart and Wurtsbaugh 2003; Teuscher 2004). However, salmonids are opportunistic and will shift their feeding to whatever is abundant, accessible, and palatable, and have sometimes have been reported with their stomachs full of unexpected prey such as snails or hornets (Jenkins *et al.* 1970; NCASI 1989; Mullan *et al.* 1992).

In general, the body of the evidence suggests that there is some ecological redundancy among aquatic stream and lake invertebrates, and if a small minority of invertebrate taxa were eliminated by chemicals at criteria concentrations, but overall remain diverse and abundant, then aquatic invertebrate overall community structure and functions, and forage value of critical habitats would likely persist. However, case-by-case consideration of the data is required because the previous assumption is tempered by the fact that aquatic insects are typically underrepresented in criteria datasets and toxicity testing in general (Mebane 2010; Brix *et al.* 2011).

Some of the anticipated effects will be to food items for juvenile salmonids, a vital component of juvenile rearing and migration habitat. Reductions in food quantity would result in limited resources to rearing and migrating fish, which can be expected to reduce population viability through increased mortality. Under-nourishment can alter juvenile salmon ability to avoid predators and select habitat within rearing drainages. Mortality can also be expected during migration, as under-nourished juveniles will not be able to withstand the rigors of migration.

Changes in species composition could have the same results. Biomass quantity is not necessarily a substitute for prey suitability, as differing prey behavior patterns and micro-habitat needs can reduce the foraging efficiency of juvenile salmonids. However, juvenile salmonids are opportunistic predators, and the loss of a minority of taxa might not be a severe indirect effect if other prey were still diverse and abundant as described above.

Effects to Other Elements of Critical Habitat. Approval of the proposed criteria may also indirectly affect safe passage conditions and access. Safe passage conditions and access to other habitats may be prevented or modified if a passage barrier exists in a section of stream because of insufficient mixing at an effluent outfall, or dilution capacity is insufficient to provide a passage corridor. To avoid these forms of adverse modification of critical habitat, the application of criteria must be protective of listed species. To determine this we evaluated if the action as proposed would provide safe passage in the manners described in Appendix F Salmonid Zone of Passage Considerations.

There appears to be little to no relation between adverse changes in water quality caused by adoption of the proposed criteria and effects to the remaining essential features of critical habitat, including: (1) Water quantity; (2) riparian vegetation; (3) instream cover/shelter; (4) water velocity; (5) floodplain connectivity; (6) water temperature; and (7) space.

2.4.2. The Effects of Expressing Metals Criteria as a function of Water Hardness

Some of the metals criteria under review in this consultation are hardness-dependent, meaning that rather than establishing a criterion as a concentration value, the criteria are defined as a mathematical equation using the hardness of the water as the independent variable. Thus, in order to evaluate the protectiveness of the hardness-dependent criteria, it was first necessary to evaluate the hardness-toxicity relations. The criteria that vary based on site-specific hardness are Cd, Cu, Cr (III), Pb, Ni, Ag, and Zn. Hardness measurements for calculating these criteria are expressed in terms of the concentration of $CaCO_3$, expressed in mg/L, required to contribute that amount of calcium plus magnesium. In the criteria equations, hardness and toxicity values and expressed as natural logarithms to simplify the math. In a general sense, these are referred to by the shorthand “ln(hardness) vs. ln(toxicity)” relations.

In the 1980s, hardness was considered a reasonable surrogate for the factors that affected toxicities of several metals. It was generally recognized that pH, alkalinity and hardness were involved in moderating the acute toxicity of metals. While it wasn’t clear which of these factors was more important, because pH, alkalinity, and hardness were usually correlated in ambient waters, it seemed reasonable to use hardness as a surrogate for other factors that might influence toxicity (Stephan *et al*. 1985). In the case of copper, dissolved organic matter or carbon (DOM or DOC) was also recognized as being important. It was assumed that DOC would be low in laboratory waters and might be high or low in ambient waters, and that hardness-based copper criteria would be sufficiently protective in waters with low DOC and conservative in waters with high DOC (EPA 1985). Most of these relations were established in acute testing, and they were assumed to hold for long-term exposures (chronic criteria). Whether that assumption is reliable was and continues to be unclear. For instance, in at least two major sets of chronic studies with metals conducted in waters with low and uniform DOC concentrations, water hardness did not appear to have a significant effect on the observed toxicity in most cases (Sauter *et al*. 1976; Chapman *et al*. 1980).

Appendix B

How to measure insignificance? Comparisons between NOECs, EC1s, and EC0s and the lower confidence limit of EC10s to estimate "insignificant effects"

Summary

To assist in our analysis, NMFS considered what toxicity test statistic best approximated a "true" no-effect concentration for evaluating risks to ESA listed species, We made a comparison of "no-observed effect concentrations" (NOECs) versus regression or distribution based methods for estimating no- or very low effects concentrations. The alternative statistics were regression- or distribution based estimates of the EC1 or EC0 (i.e., concentrations causing adverse effects to 1% or 0% of a test population), and the lower 95th percentile confidence limit of the concentration affecting 10% of the test population (LCL- EC10), which is a statistic used in human health risk assessment for determining benchmark doses of materials that present low increased risk (EPA, 2000a), Our conclusion was that if the data sets had a gradient of effects that would allow calculation of an EC0, the EC0 would be the preferred, best estimate of no-effect value from a toxicity test. If data were insufficient to calculate an EC0, the NOEC may be the best appropriate statistic.

The problem

In evaluations of the risks of chemicals to aquatic species listed as threatened or endangered, the statistical interpretation of toxicity testing has become an issue. Classically, the interpretation of chronic or sublethal tests has involved the use of statistical hypothesis testing, the results of which are commonly reported as "no-observed effect concentration" (NOEC) or "lowest-observed effect concentration" (LOEC). Definitions vary, but for this analysis the LOEC will be considered the lowest concentration for which there is a 95% probability that the biological response of interest (survival, growth, fecundity, etc.) is different from the control response. Similarly, the NOEC is considered the next lowest treatment. It has been assumed that somewhere between the NOEC and LOEC lies a maximum acceptable toxicant concentration (MATC) that represents a "true" but unknown threshold for unacceptable effects. In practice, the MATC concentration is estimated as a simple geometric mean between the NOEC and LOEC (Gelber *et al.* 1995). This is the value usually used in EPA criteria documents to estimate "safe" concentrations from a chronic toxicity test, although the term "MATC" is avoided in the Guidelines and instead the statistic is called a "chronic value" for a test. MATCs in turn are averaged to obtain species mean chronic values, and ultimately to set chronic criteria values.

The EPA criteria approach seems to conflict with concepts for evaluating risk to listed species because the EPA approach of averaging NOECs and LOECs assumes that aquatic communities are resilient to, or can recover from, some low-level of adverse effects. In contrast, if a species was listed as threatened or endangered, it is assumed to have substantially less resiliency than general aquatic communities. Therefore, in interpreting toxicity test data, a statistic that by definition includes some uncertain but probably low level of adverse effect such as the EPA "chronic value" is inappropriate as a statistic of effects on listed species that are expected to be discountable or insignificant. In the ESA Consultation Handbook for evaluating effects of actions to listed species, states that*" 'insignificant effects' relate to the size of the impact and should never reach the scale where take occurs. Discountable effects are those extremely unlikely to occur. Based on best judgment, a person would not: (1) be able to meaningfully measure, detect, or evaluate insignificant effects; or (2) expect discountable effects to occur." (USFWS and NMFS 1998).* Thus a meaningful measurement of low-effects from a toxicity test such as an EC10 or EC5 is inherently in conflict with a definition that requires "insignificant" effects to be unmeasurable.

An obvious substitute for use in ESA consultations is the NOEC, and indeed that is the default statistic selected in EPA's methodology for conducting biological evaluations of aquatic life criteria (EPA 2003). However, in recent years the concept of the NOEC has been battered in the ecotoxicology literature. The three complaints relate to the common design of toxicity experiments which usually involve a series of about five treatment concentrations plus a control, each replicated about three times. Complaint #1 is that a NOEC has to be one of the concentrations tested, so its precision is dependent on the number and spacing of treatment concentrations. So for example, if the unknown "true" no-effect concentration is 1.8 µg/L a test series of 1.0, 1.2, 1.6, 2.0, …. will give a more precise NOEC estimate than a series of 1, 2, 4, 8, 16, ….(1.6 vs. 1.0 µg/L). Complaint #2 is that for the low levels of replication used (often 3), the minimum statistically detectable effect level can vary widely, easily from 5 to about 40% for endpoints with low or high variability (e.g. growth in fish (low) or fecundity in invertebrates (highly variable). The NOEC statistic by itself gives no insight into whether a "significant" effect is biologically trivial or whether an effect is biologically serious but too variable to be significant at the arbitrary limit that no more than a 5% risk of being wrong is acceptable (acceptable to the evaluators, not whether it is acceptable to the organism). Complaint #3 is related in that the NOEC-LOEC approach is solely focused on the "Type I" error, or the risk of declaring an adverse effect when the observed effects occurred solely by chance, with no or little regard for Type II error, the risk of failing to detect an adverse effect that was really present but the test had insufficient power to detect it. Type II error rates may be quite high in ecotoxicological studies that fail to detect effects as "significant" at the 5% Type I error rate (Stephan and Rogers, 1985; Laskowski, 1995; Moore and Caux, 1997; Crane and Newman, 2000; McGarvey, 2007; Newman, 2008; Brosi and Bilber, 2009).

An alternative often put forth to the NOEC-LOEC approach is regression or distribution based techniques that fit an effects curve to the observed data, and then any point along that curve can be used to estimate effects at a given concentration. This regression or distribution based approach is the most common technique for defining LC_{50}s in acute data but obviously other effect concentrations percentiles (ECp) besides the 50^{th} percentile could be of interest. The catch in this approach is that it is up to the assessor to independently determine what level of effect is "important." Choices of what level of effect is "important" have either been made subjectively or by comparisons of ECp values back to NOECs and LOECs. For example, in the interpretation of EPA's chronic whole-effluent toxicity (WET) tests, NOECs are assumed to be equivalent to an EC25. The conclusion that a 25% adverse effect in a biologically important endpoint therefore represents a no-observed effect concentration was supported by a citation to an analysis of 23 pooled chronic WET test results for red algae, sheepshead minnows, sea urchins, *Ceriodaphnia*, and fathead minnows in which NOECs were more frequently similar to EC20s (EPA, 1991, p. 27). No reason was given why the EC25 was endorsed over the EC20, since the analysis supported the use of the EC20, but regardless the EC25 is often the trigger statistic in WET tests.

Subsequent analyses have also shown that NOECs are usually higher than point estimates of low toxic effects such as the EC10 (Moore and Caux, 1997; Crane and Newman, 2000). In an analysis limited to the effects of cadmium, NMFS found that the typical expected adverse effect associated with MATC was often about 20-30% with invertebrates and about 10-15% for fish (Mebane, 2006). However, using ECx values that correspond with a NOEC or MATC to select "x" as a suitable replacement for the unsuitable NOEC falls into circular reasoning. A counterpoint could be made that comparisons of ECps and NOECs to support an ECp value to replace NOECs is a tautology. Instead of matching statistics, biological arguments could be made for assuming different "acceptable" ECp values based upon patterns of variability of the same endpoints in natural populations, life history strategies, projecting effects in population models, and field studies relating year class survival to size differences. No comprehensive analysis along these lines is known to have been published.

Mebane and Arthaud (2010) gave an example of what effect-statistics could be related to population extinction risks or recovery trajectories for a headwaters threatened Chinook salmon population. In this population, Marsh Creek in the upper Salmon River, Idaho, survival of juvenile migrants is strongly related to the size of the fish. A size reduction of 4% as length, i.e., an EC04, was associated with survival reductions ranging from 12 – 38% for different migrant groups from a trap near the headwaters to the first dam encountered downstream. In the toxicity tests with Chinook salmon and rainbow trout that were analyzed for the study, a 4% reduction in length corresponded with about a 12% reduction in weight. When the survival reductions associated with a length EC04 were extrapolated through a population model to changes in extinction risk or recovery time, little difference in extinction risk was projected but an appreciable delay in recovery was projected. This indicates that at least for the length endpoint in chronic fish toxicity tests, the statistical threshold for important adverse effects may not be much higher than statistics such as an EC0 or EC01. Yet for the commonly used weight endpoint in chronic fish toxicity tests, the statistical threshold for important adverse effects would be higher, around the EC10. Presumably, if endpoints are more variable, such as the number of eggs produced per female (fecundity), then a higher ECp value (e.g. EC20) might be appropriate. While the relevance of this example to other species or even different populations of Chinook salmon is not known, it does at least serve as one example of a basis to judge the importance of an ECp value without relying on circular comparisons back to other statistics.

Comparisons between statistics

For this exercise, NMFS evaluated data from a variety of available toxicity tests results that were available in the syntax required by the statistics models. While such data are not comprehensive or necessarily definitive, they are preferable to many journal articles because the latter are sometimes too summarized to make any subsequent analysis of. We selected the examples to illustrate a variety of response patterns ranging from classic, concentration-responses to test results that are difficult to interpret.

NMFS used either reported NOECs or those that could be estimated using Dunnett's test. ECp values were estimated for growth and reproduction using a distribution analysis for survival data (respondents are either alive or dead) or nonlinear regression for more or less continuous data (growth or fecundity measurements). For each type of analysis, a choice of underlying distributions of the populations must be assumed.

(1) Gaussian (Normal) Distribution: This is based on the familiar "bell curve" or gaussian distribution. This produces a sigmoidal toxicity relationship with infinite tails, and is equivalent to prohibit analysis.

(2) Triangular Distribution: This produces a sigmoidal toxicity relationship similar to the gaussian distribution, but with a finite threshold exposure below which responses are zero and a finite exposure above which all organisms are affected. It is also referred to as a "sigmoid threshold" (Erickson 2008).

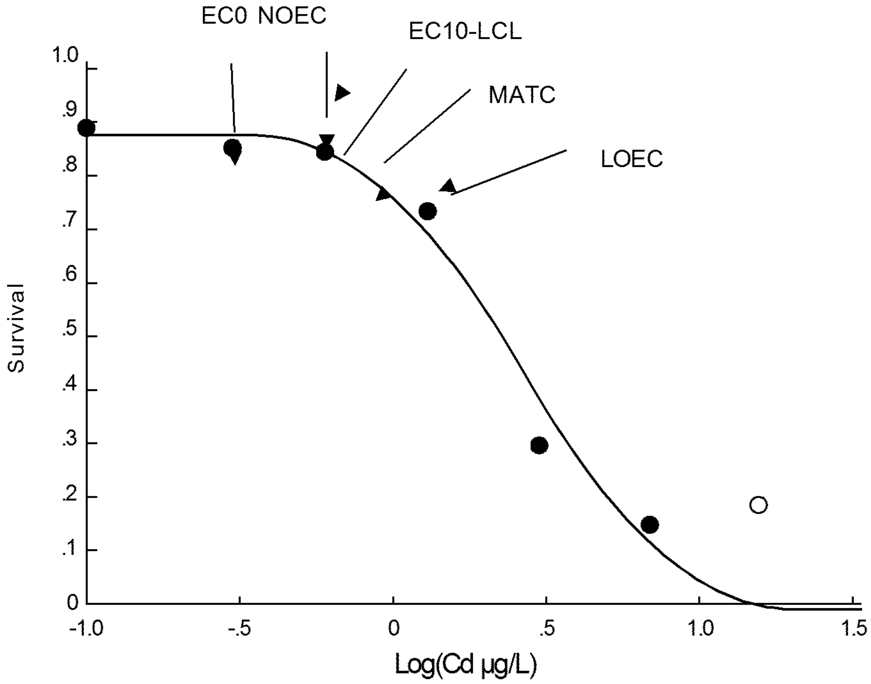
(3) Uniform (Rectangular) Distribution: This produces a piecewise-linear toxicity relationship, for which there is a finite lower and upper exposure limit like the triangular distribution, but for which the decline in response between these limits is linear rather than sigmoidal. Similar analyses have been called "jackknife distributions" in the literature because of its shape.

The assumed statistical distribution and behavior of the data in the tails of the distribution are usually of little consequence when one is trying to estimate the middle of the distribution (LC_{50}). However, when one is trying to estimate no-effects data, these estimates are at the extreme tails of the distribution, and the shape of the tails and the behavior of the models become more important. In the Gaussian, normal distribution, an EC0 can never be achieved because the tails are infinite; in other words some rare organisms are assumed to be infinitely resistant and some sensitive to infinitesimal exposures. Because that assumption is not plausible for ecotoxicology data, methods have been developed using discrete distributions with definite ends, i.e. no organism is infinitesimally sensitive, and an EC0 can be calculated.

NMFS calculations used a beta version of the Toxicity Response Analysis Program, under development EPA's National Health and Environmental Research Laboratory, Mid-Continent Ecological Division (Erickson 2008).

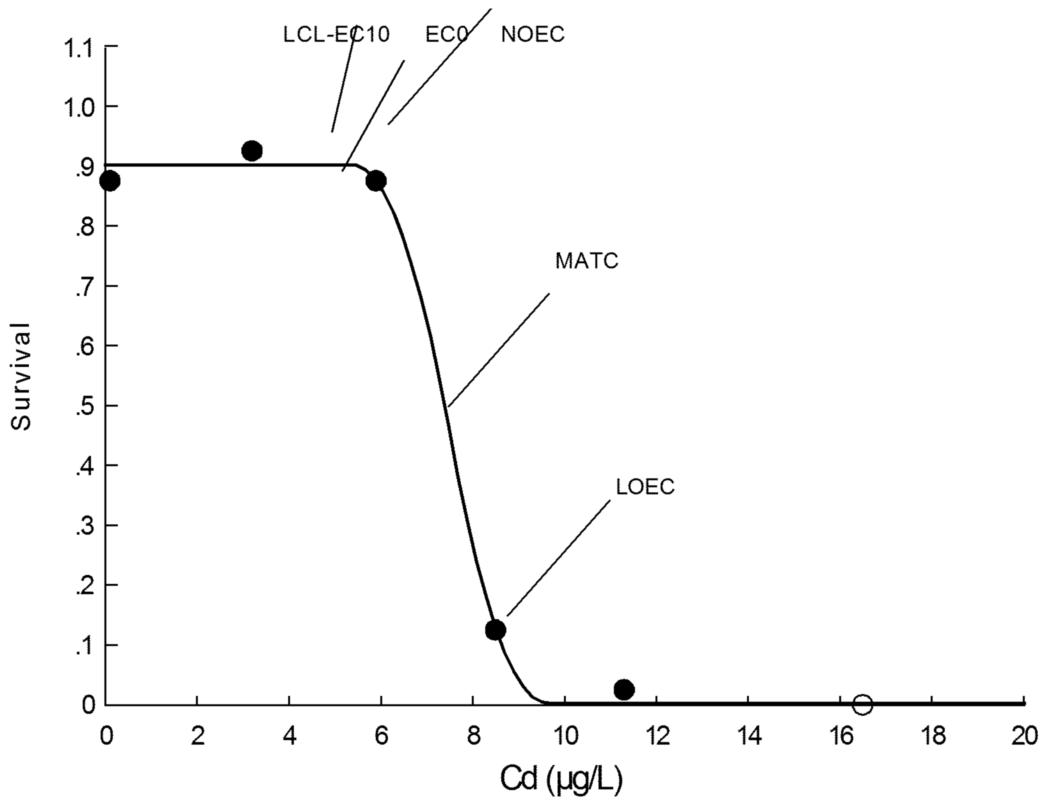
Examples:

Example 1. Rainbow trout 53-day survival with cadmium, using the sigmoid threshold model based upon an assumed triangular distribution. Open circles indicate data points that were excluded from the regression



ECp	ECp est	95 LCL	95% UCL
10	0.85	0.62	1.17
0	0.35	0.21	0.59

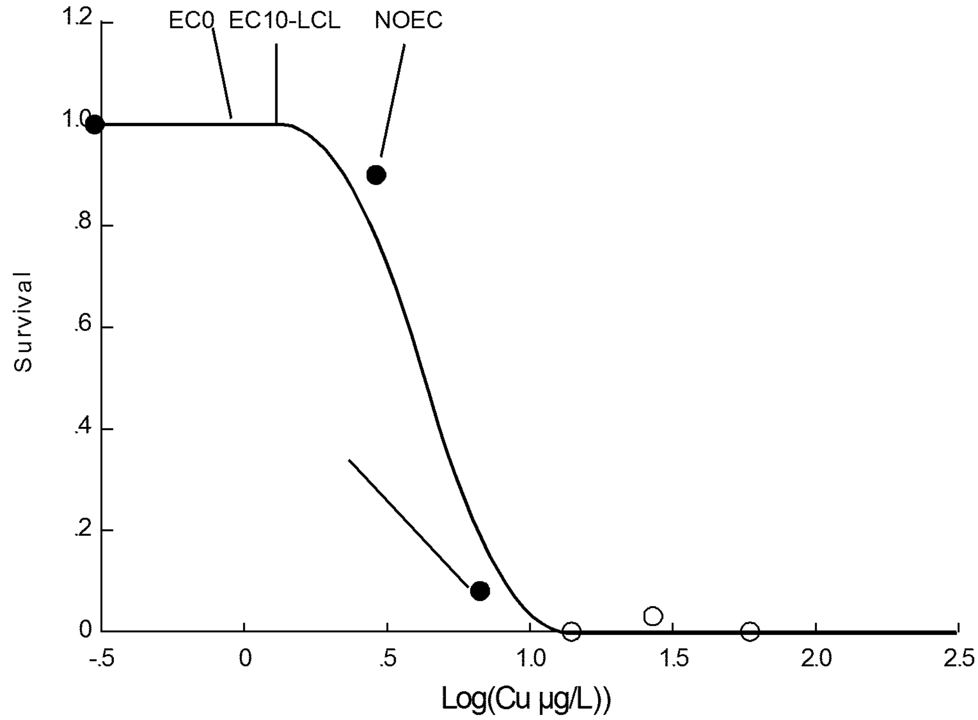
Example 1 was selected to illustrate the classic ski jump curve shape, where the initial part of the curve from the control out to the 2nd treatment shows a slight decline, followed by a steep drop in the center region of the curve where intermediate effects occur, followed by a flattening out of the slope at the bottom as almost all animals are predicted to be killed (Mebane *et al.*, 2008).



Example 2. Fountain darter, 7-day survival with Cd, sigmoid threshold, showing a very steep curve that results with (nearly) all-or-nothing responses. In this case, all of the "nearly-no-effect" estimators give similar values.

ECp	ECp est	95 LCL	95% UCL
10	6.33	5.06	7.59
0	5.38	2.84	7.92

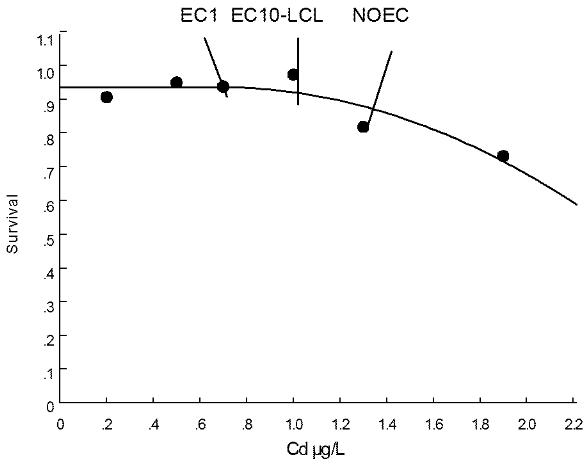
(Castillo and Longley, 2001)

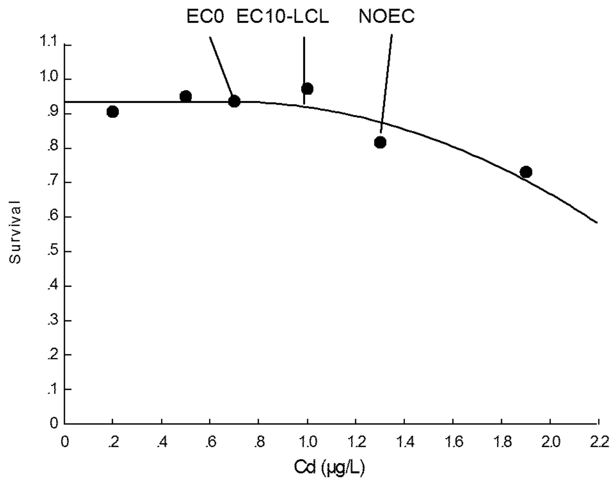


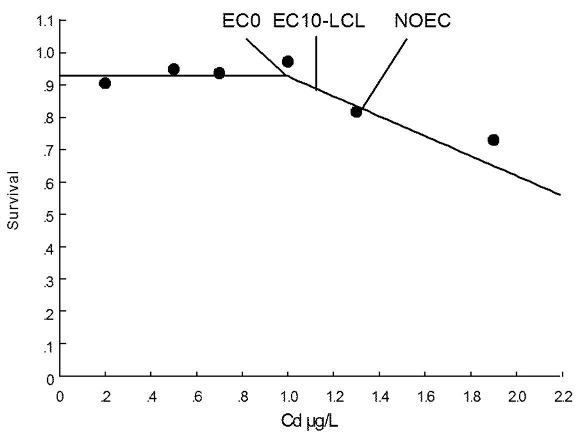
Example 3. Mottled sculpin, 14-day survival with copper (Besser and others, 2007). As with example 2, these data had inadequate partial responses resulting in an uncertain fit between the control and treatment 1, the NOEC. Even so, ECp estimates are reasonable and confidence limits are not large. These type of data are often encountered working with listed species or other poorly tested species for which investigators have little idea in advance what exposure change to test.

ECp	ECp est	95 LCL	95% UCL
10	2.255	1.841	2.762
5	1.934	1.516	2.466
0	1.334	0.924	1.925

LOEC

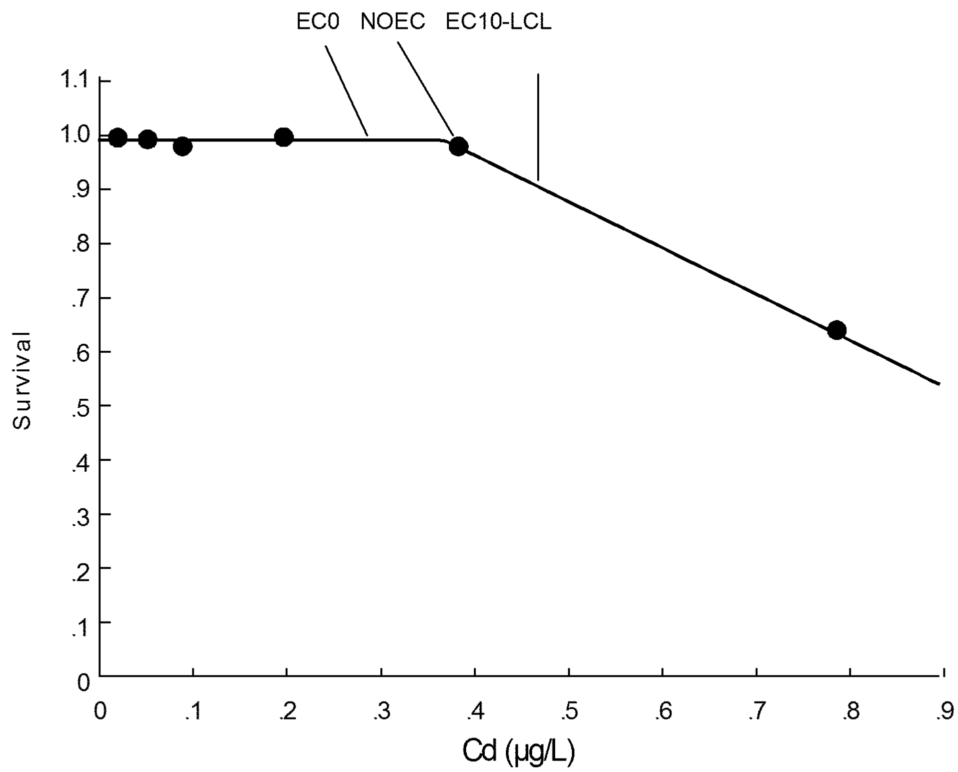




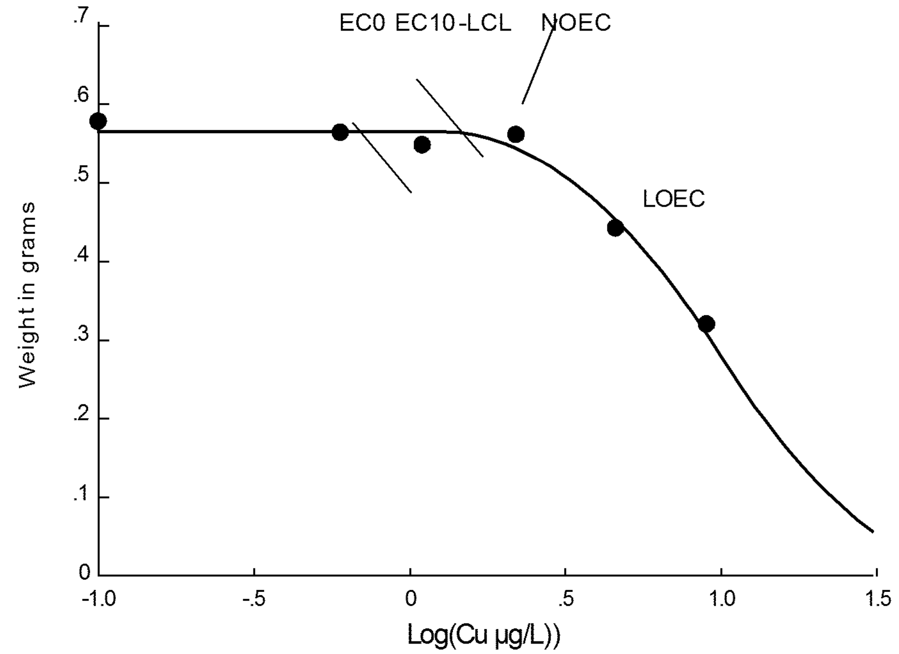


Example 4. Chinook salmon 120 day survival with Cd (Chapman, 1982), illustrating differing ECp estimates resulting from different statistical models. Note that EC0s are conceptually impossible using the normal distribution, but the EC1 in the top figure is close to the EC0 in the middle figure using the triangular distribution. In this example, the linear model (bottom) does the best job of finding the no-effect estimate (visually, treatment 3, the 4th point from the left). Despite the very different underlying models, all ECp estimates were similar in this example.

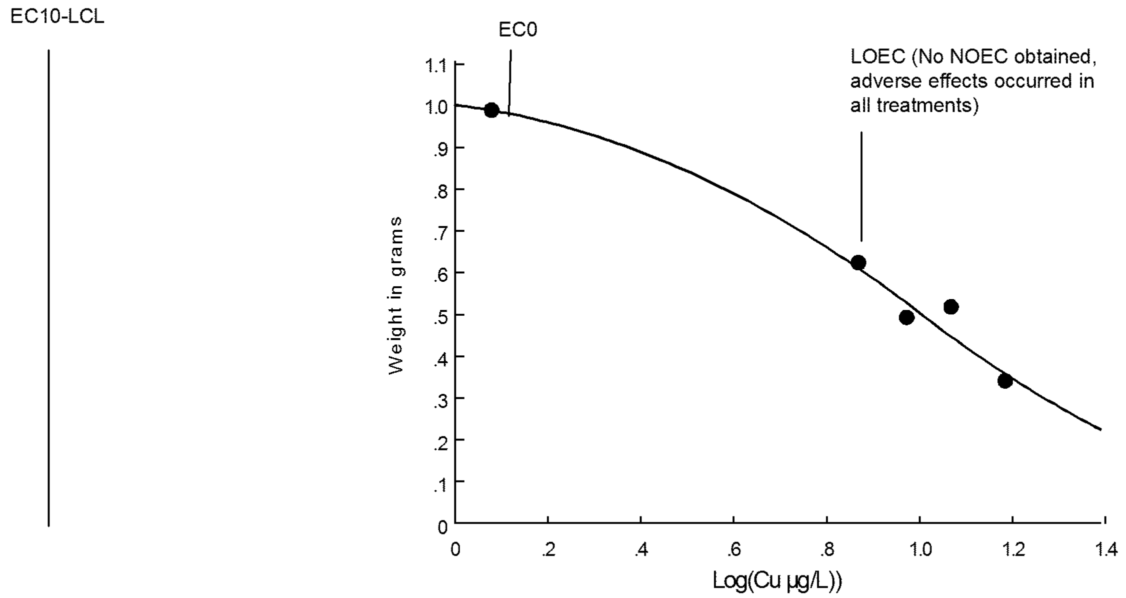
Chinook and Cd ECp values				Bull trout and Cd ECp values:			
Gaussian	ECp est	95 LCL	95% UCL	ECp	ECp est	95 LCL	95% UCL
20	1.802	1.541	2.063				
10	1.480	1.042	1.919				
5	1.215	0.503	1.927				
1	0.717	-0.789	2.223				
Triangular							
20	1.792	1.529	2.056				
10	1.466	1.144	1.788	10	0.555	0.477	0.633
5	1.236	0.764	1.707	5	0.479	0.379	0.578
0	0.679	-0.454	1.811	0	0.294	0.117	0.471
Rectagular							
20	1.609	1.366	1.852	20	0.60496	0.58075	
10	1.304	1.114	1.495	10	0.48715	0.46026	
5	1.152	0.953	1.352	5	0.42824	0.39815	
0	1.000	0.763	1.237	0	0.36933	0.33522	



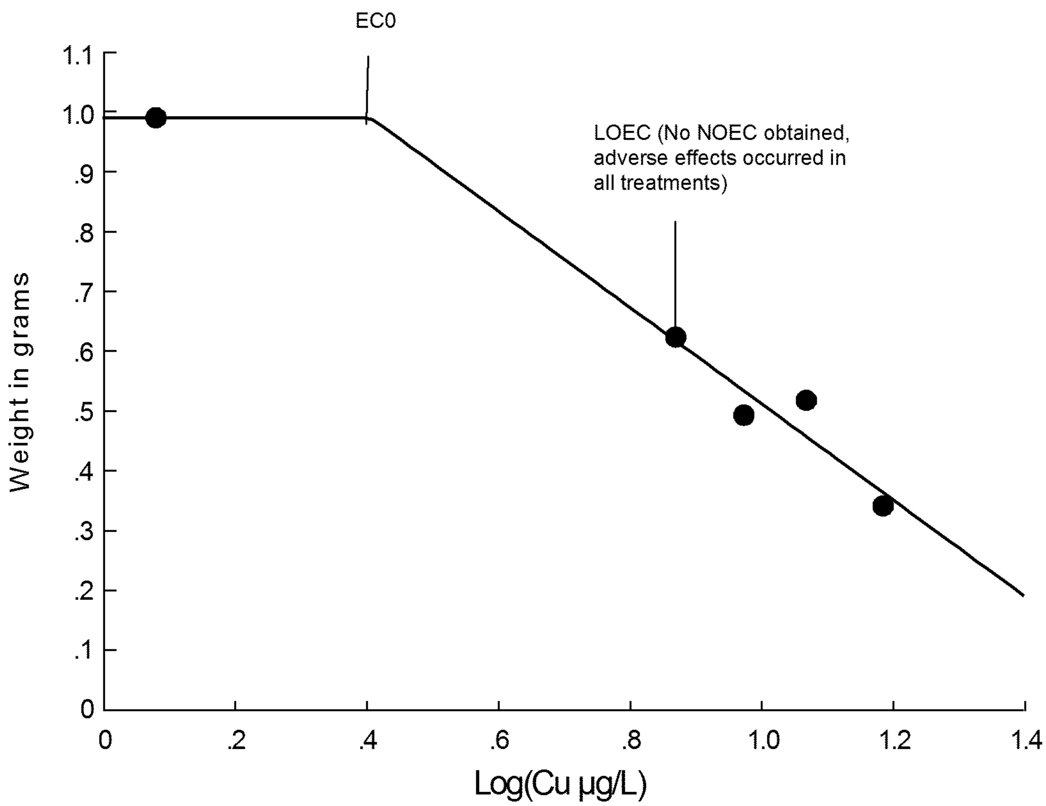
Example 5. Bull trout, 55-day survival with Cd (Hansen and others, 2002c). Here the NOEC is lower than the LCL-EC10. Similar to the Chinook salmon and Cd example, these data would give an inadequate and highly unreliable response for an LC50. However, with chronic testing the interest is in the low-effect part of the curve.



Example 6. Growth of rainbow trout after 60-days Cu exposure (Marr and others, 1996). This data set is nicely balanced with 3 nearly no-effect treatments and 2 treatments above a clearly defined effects threshold.



Example 7. Growth of Chinook salmon after 120-days Cu exposure, sigmoid threshold model (Chapman, 1982). This data set presents uncertain EC0 values because adverse effects occurred in all tested treatments. The LCL-EC10 is less than zero which is clearly impossible and using the sigmoid model, the EC0 falls close to the control. There is no NOEC, although in some data compilations the "less than" for this treatment was lost in translation and the NOEC or chronic value has been treated as 7.4 µg/L rather than < 7.4 µg/L. This mistake results in a 40% reduction in growth being treated as a low- or no-effect.



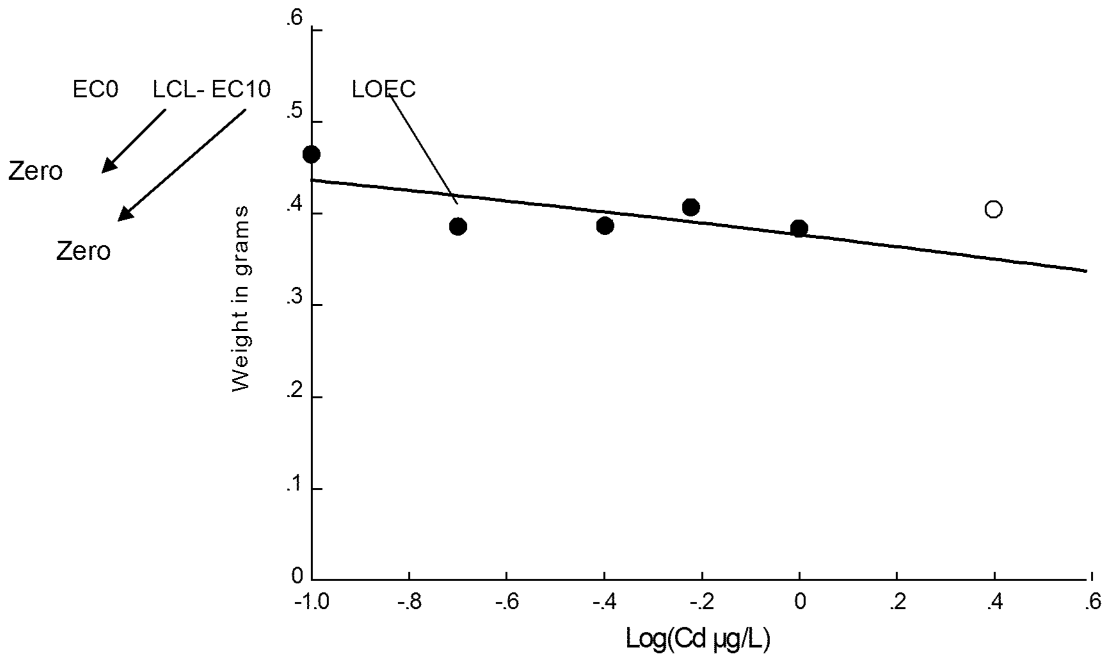
Example 8. Growth of Chinook salmon after 120-days Cu exposure, piecewise linear response (Chapman, 1982). Curves do not always give better fits; here it is more plausible that the onset of adverse effects occurs at a higher copper concentration than the controls. However, in data sets such as this, the interpolation between the control and first treatment data set is so large that the shape of the curve and thus the response is less a statistical question than a professional judgment about what seems most plausible.

Chinook growth (sigmoid threshold)

ECp	ECp est	95 LCL	95% UCL
10	2.215	0.026	185.270
1	0.954	0.000	12238.000
0	0.646	0.000	652500.000

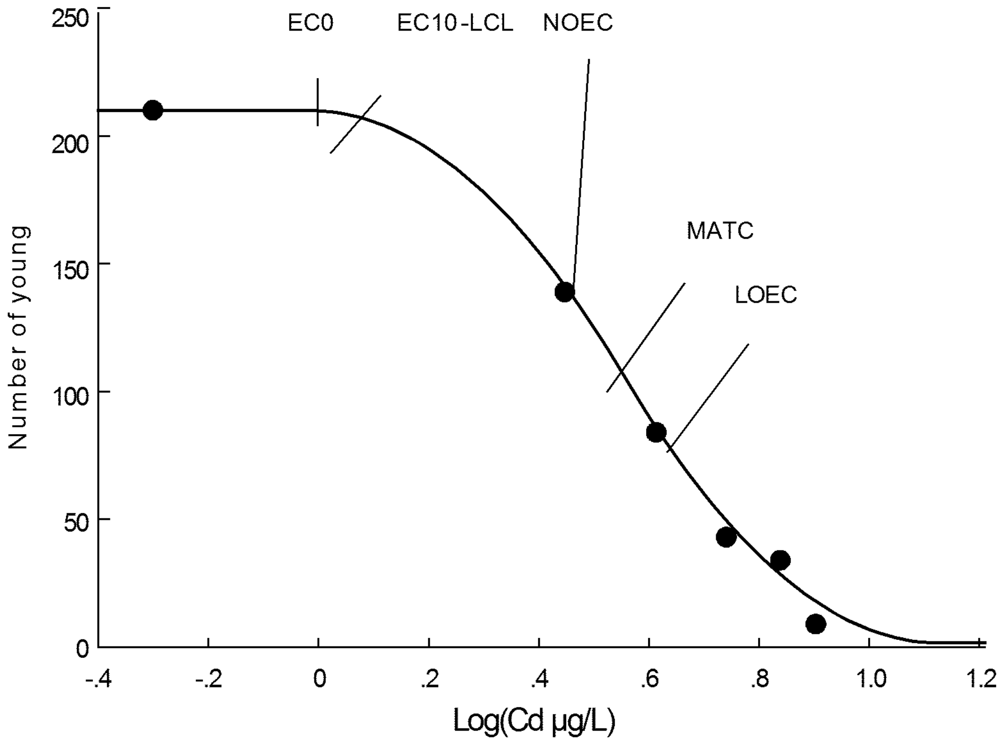
Chinook growth (piecewise linear)

10	3.386	0.699	16.399
1	2.623	0.396	17.354
0	2.550	0.372	17.468



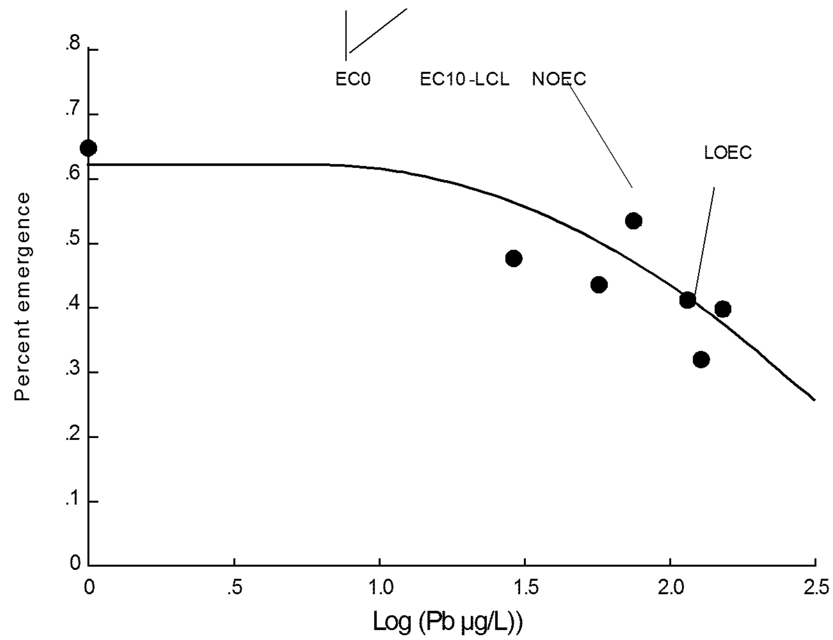
Example 9. Rainbow trout growth after 62-d exposure to Cd (Mebane *et al.* 2008). This example is similar to the Chinook salmon and Cu example in that statistically significant effects were observed in all treatments and no NOEC could be obtained. Further, because no monotonically decreasing concentration response was observed, the curve was almost flat and ECp values are meaningless (numerous errors and warnings were overridden to create this example). In this example, statistics of any type offer little help in interpreting the data.

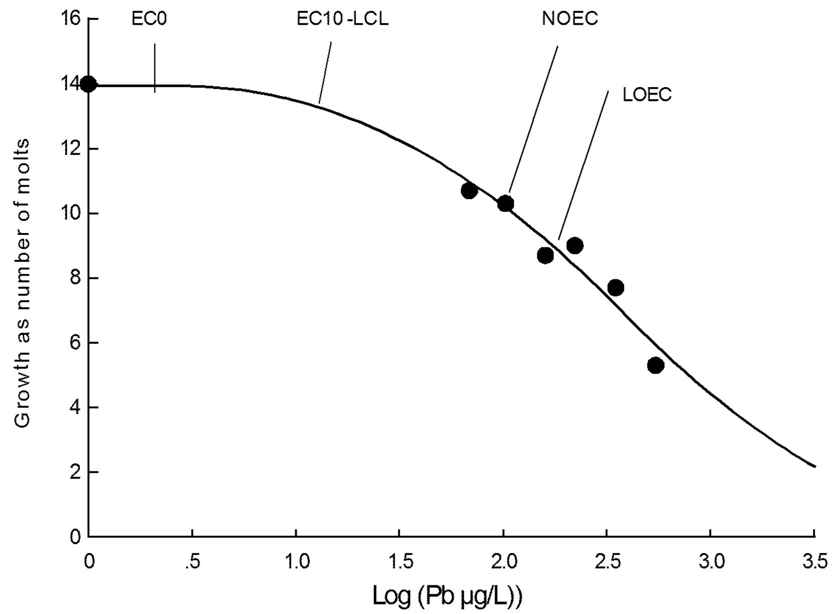
ECp	ECp est	95 LCL	95% UCL
50	16.61600	0	Infinity
20	0.02234	0	Infinity
10	0.00080	0	Infinity
5	0.00008	0	Infinity
0	0.00000	0	Infinity



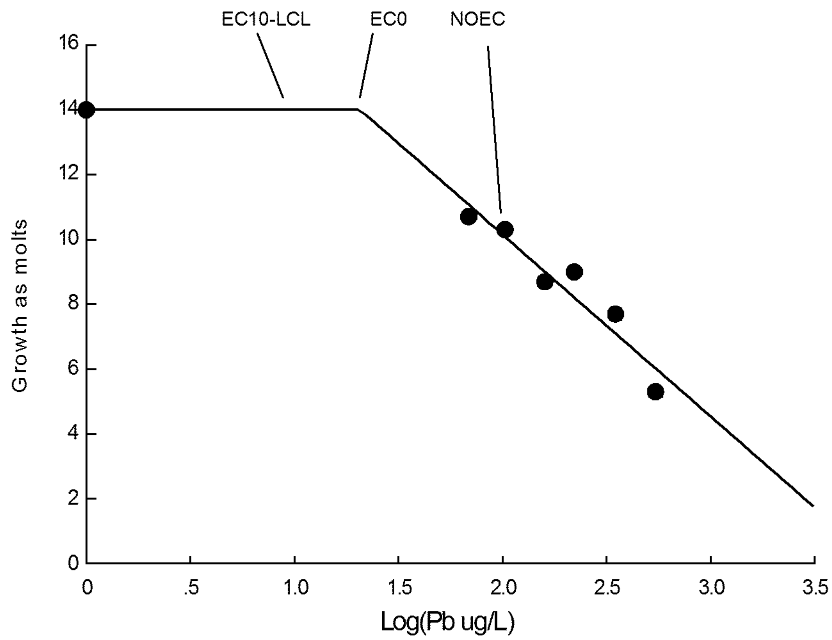
Example 10. Reproduction of *Ceriodaphnia dubia* after 7-d exposure to Cd (Castillo and Longley, 2001). In this test, the NOEC reported by the authors corresponded to about a 35% reduction in reproduction, and greater than a 50% reduction for the MATC .

ECp	ECp est	95 LCL	95% UCL
Sigmoid	2.800		
25	2.414	1.970	2.957
10	1.716	1.239	2.375
1	1.148	0.577	2.282
0	0.953	0.482	1.887





Example 11. Emergence of midge (*Chironomus tentans*) larvae following 21-days exposure to Pb (Top); Mayfly (*Baetis tricaudatus*) molting during 10-days exposure to Pb (Mebane *et al.* 2008). Examples of less than ideal datasets that can arise from testing of non-standard organisms or tests conducted in environmentally realistic but noisy experiments (these were streamside tests). The shape of the curves in both datasets suggest an onset of effects below the lowest concentration tested. This suggests both that NOECs may not be conservative and that low ECp values are uncertain.



Example 12. (Continued) Same mayfly (*Baetis tricaudatus*) as above, but using a piecewise linear or jackknife distribution. As with the case of copper and Chinook salmon growth, assuming a curved distribution would cause the EC0 estimates to be near the control. If that were to be considered implausible, the jackknife "curve" provides a higher "no-effect" value that statistically is equally valid.

C. tentans, Pb Emergence, logistic			
ECp	ECp est	95 LCL	95% UCL
10	30.697	5.793	162.670
5	19.039	1.962	184.720
0	6.009	0.067	540.460

Mayfly, Baetis tricaudatus - logistic				**Mayfly, Baetis tricaudatus – "Jackknife**		
ECp	ECp est	95 LCL	95% UCL	ECp est	95% UCL	95% UCL
20	63.159	25.394	157.090	65.972	28.888	150.66
10	25.713	6.721	98.375	37.103	13.112	104.99
5	13.620	2.514	73.806	27.825	8.7638	88.342
0	2.937	0.206	41.873	20.867	5.8379	74.585

Conclusions

In most of these comparisons, the rank order of the "effects" concentrations were EC0< LCL-EC10 <NOEC. Of the statistics examined, the LCL-EC10s seems particularly suspect. Generally, LCL-EC10 estimates were close to EC0 or EC1 values, however, in all cases where reasonable LCL-EC10 estimates could be obtained, so could EC1 or EC0 values. Confidence intervals on very low effect estimates are large, but at least for EC1 or EC0 values, confidence intervals can be calculated. No confidence limits can be calculated on a confidence limits, and there is no logical reason why the LCL-EC10 is a better estimate of an EC1 or EC0 than would be the EC1 or EC0 themselves. In sum, no empirical or theoretical reason for using LCL-EC10 statistic could be envisioned.

In most instances, the differences between the NOECs, LCL- EC10s, and EC0s were small. This suggests that given the magnitude of uncertainty involved in other aspects of evaluating risks to listed species such as extrapolating effects between species, and extrapolating acute-to-chronic effects, the choice of which statistic used to estimate "no-effect" for a given test response may be of less importance. Some datasets were less than ideal for the statistical models. For most datasets, estimates of these extreme statistics seemed reasonable, based on the datasets from which they were derived. Confidence limits were very large, but the estimates themselves seemed reasonable. Some ECp analyses were uncertain, most commonly because of inadequate partial effects resulting in uncertainty in the shape of the response curve. In other tests, adverse effects resulted in all treatments, so no NOEC could be determined. Differences in results obtained using different assumed statistical distributions (normal, triangular, rectangular) were small.

The results of NMFS' analysis suggest that for initial screening of large databases for chronic effects concentrations to compare with criteria values, any of the NOEC, LCL-EC10, EC1, or EC0 statistics could be useful, and the choice of which statistic to use will probably depend on which is most available. However, in instances where the test is influential in the assessment, a more careful review of the original research might enable the assessor to make a more informed judgment of whether the test indicates reassurance of the lack of effects or indicates that adverse effects are likely. These judgments cannot always follow rote statistical analyses.

References for Appendix B

Besser, J.M., Mebane, C.A., Mount, D.R., Ivey, C.D., Kunz, J.L., Greer, E.I., May, T.W., and Ingersoll, C.G., 2007, Relative sensitivity of mottled sculpins (*Cottus bairdi*) and rainbow trout (*Oncorhynchus mykiss*) to toxicity of metals associated with mining activities: Environmental Toxicology and Chemistry, v. 26, no. 8, p. 1657–1665 http://dx.doi.org/10.1897/06-571R.1

Brosi, B.J., and Bilber, E.G., 2009, Statistical inference, type II error, and decision making under the US Endangered Species Act: Frontiers in Ecology and the Environment, v. 7, no. 9, p. 487-494. http://dx.doi.org/doi: 10.1890/080003